THE UNITED STATES DISTRICT COURT
FOR THE NORTHERN DISTRICT OF TEXAS
DALLAS DIVISION

FEDERAL TRADE COMMISSION,

Plaintiff,

vs.

MATCH GROUP, INC., a corporation, and MATCH GROUP, LLC, formerly known as MATCH.COM, LLC, a limited liability company,

Defendants.

Case No. 3:19-cv-02281-K

APPENDIX IN SUPPORT OF DEFENDANTS REPLY IN SUPPORT OF THEIR RULE 702 MOTION TO EXCLUDE DR. JENNIFER KING'S TESTIMONY

Defendants, by and through their counsel, submit this Appendix in support of their Reply in support of their Rule 702 Motion to Exclude Dr. Jennifer King's Testimony.

No.	Description	App. Page(s)[1]
1.	Dr. Jennifer King's Notes, with filename "Match brainstorming"	App. 446-462
2.	Jennifer King Deposition Transcript (July 27, 2023) Excerpts	App. 463-467
3.	*FTC v. Commerce Planet, Inc.*, Dkt. 125, Defendant Charles Gugliuzza's Motion *in Limine* for Order Excluding Expert Testimony of Jennifer King Pursuant to *Daubert v. Merrell Dow Pharmaceuticals, Inc.*; Memorandum of Points and Authorities In Support Thereof (Apr. 29, 2011)	App. 468-483
4.	*FTC v. Commerce Planet, Inc.*, Dkt. 293, Reporter's Transcript of Proceedings (Feb. 7, 2012)	App. 484-489
5.	Article: How to Run an Effective Heuristic Evaluation	App. 490-497
6.	Article: An Empirical Study of Usability Testing: Heuristic Evaluation vs. User Testing	App. 498-502

[1] Appendix pagination is continuous from the appendix filed with Defendants' Motion. *See* Dkt. 209. References in the brief to "Mot. App." are to Dkt. 209. References in the brief to "Resp. App." are to the appendix filed with the FTC's Response, at Dkt. 220-3. References in the brief to "App." are to this appendix, filed with Defendants' reply.

No.	Description	App. Page(s)[1]
7.	Article: Finding Usability Problems through Heuristic Evaluation	App. 503-510
8.	Article: Heuristic Evaluation of User Interfaces	App. 511-518
9.	Article: Extracting Usability and User Experiences Information from Online User Reviews	App. 520-528
10.	Article: Heuristic Evaluation vs. User Testing	App. 529-532
11.	Article: Customer Complaint Behaviour and Companies' Recovery Initiatives: The Case of Hello Peter Website	App. 533-555

Dated: October 16, 2023

/s/ Angela C. Zambrano

Angela C. Zambrano
State Bar No. 24003157
angela.zambrano@sidley.com
Chelsea A. Priest
State Bar No. 24102375
cpriest@sidley.com
Tayler G. Bragg
State Bar No. 24109943
tbragg@sidley.com
SIDLEY AUSTIN LLP
2021 McKinney Avenue, Suite 2000
Dallas, TX 75201
Telephone: 214-981-3300
Fax: 214-981-3400

Chad S. Hummel (admitted *pro hac vice*)
chummel@sidley.com
SIDLEY AUSTIN LLP
1999 Avenue of the Stars, 17th Floor
Los Angeles, CA 90067
Telephone: 310-595-9500
Fax: 310-595-9501

Benjamin M. Mundel (admitted *pro hac vice*)
bmundel@sidley.com
SIDLEY AUSTIN LLP
1501 K Street, N.W.
Washington, DC 20005
Telephone: 202-736-8000
Fax: 202-736-8711

Attorneys for Match Group, Inc. and Match Group, LLC

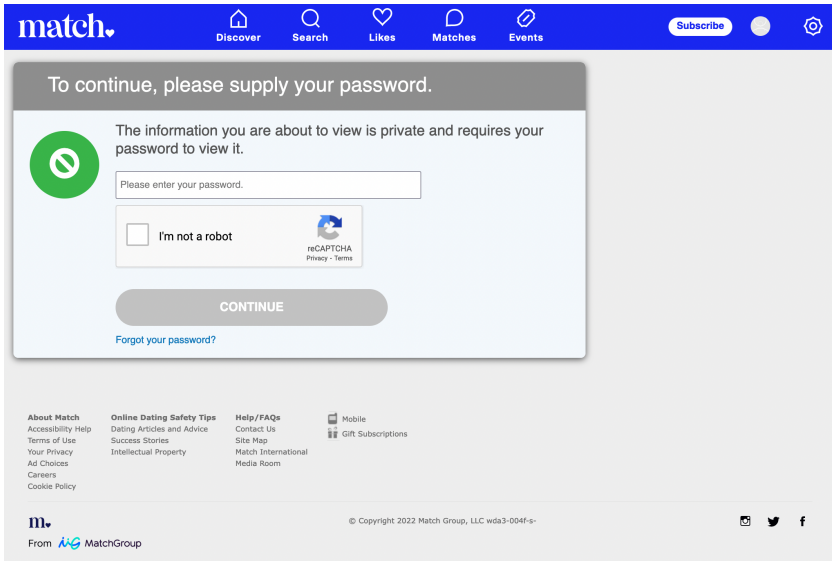
CERTIFICATE OF SERVICE

I hereby certify that on October 16, 2023, I caused a true and correct copy of the above and foregoing document, to be served on all counsel of record in accordance with the Federal Rules of Civil Procedure and this Court's CM/ECF filing system.

/s/ Angela C. Zambrano
Angela C. Zambrano

Cancellation flow:

Adds unnecessary friction
Requires a password to manage subscription
Even on a free account, you get this password barrier:



BUT! Clicking manage subscription appears to be the only thing that brings up the password window. Clicking any other option (including delete!) doesn't give you a password request. (Instead, with Delete you get a bunch of guiltshaming steps.)

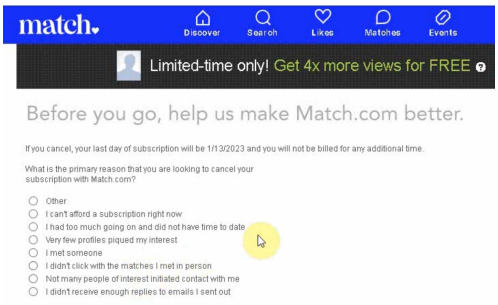
TODO: map out flow w/no extra steps
TODO: map out flow including the upsell step
TODO: calculate max number of steps in existing flow
TODO: find references for quantifying friction/when adding additional steps or hurdles to an online flow translates to lost customers

- Nielsen/Norman
- Surveys - things you have to fill out
- Radio buttons - effect of that
- Is a cancellation flow the best place to offer a survey? (could we standardize the process for post-cancellation surveys, for example?) - forcing users to do it as part of the cancellation process is inappropriate
- Password friction
- Review Match.com cancellation steps: https://help.match.com/hc/en-us/articles/6077124196891-Cancelling - note that they don't provide detailed steps for canceling an account, but have very detailed steps for canceling additional features
- Sign-up for Match 3 day free trial? https://www.match.com/cpx/en-us/landing/search/208264-free-trial/

Survey questions are optional

Which dark patterns are at play here?

1. **Obstruction** (Making a process more difficult than it needs to be, with the intent of dissuading certain action(s).) – specifically, roach motel (can't cancel). The insertion of excessive/unneeded steps into the cancellation flow is obstruction. (https://darkpatterns.uxp2.com/ & https://webtransparency.cs.princeton.edu/dark-patterns/)
2. **Forced action** (Requiring the user to perform a certain action to access (or continue to access) certain functionality.) – the survey questions (despite being "optional") are a form of forced action - giving the user the impression that they are required. The user would have to take the time to experiment in order to ascertain that they aren't necessary. Password prompt might be classifiable as forced action.
3. Misdirection? (Using visuals, language, or emotion to steer users toward or away from making a particular choice.) – this might be a question of analyzing language, placement of links, etc. carefully
    a. The language used might encourage users to fill out the survey, even though it's optional. Exs: "Tell us more." "Before you go, help us make Match.com better." These statements lead users to believe they *must* divulge their experience before canceling. A more honest way for Match to present itself would be "Would you like to take a survey?" making it clear that it is optional.



    b. In some cases, I can see the "1000 characters remaining field" being a deterrent. Users might click back to the survey questions, change their answers in hopes of not getting this open ended text field.

    c. "Lose the benefits," "you risk losing," "you won't know." Canceling the service is framed as a **loss** rather than a plain option. A more neutral way might be, "If you cancel now, you will lose access to: who's viewed your profile, email communications, our current monthly rate." also "resign" is a loaded word!

JP: These are some of the initial dps identified from the report and documents. If any are beyond the scope of what we're mapping, let me know and I'll strike it.

- Bait and Switch: Nonsubscribers receive emails indicating a Match.com user has expressed interest in him or her ('bait'). Upon buying a subscription to Match.com new subscribers learn the account that reached out to them was fraudulent or 'unavailable' for viewing ('switch').
- Trick Questions: The Guarantee Program Subscription indicates that users are eligible for a Guarantee Extension (free six-month subscription) if they have not met their "special someone." In obtaining the extension, users are asked, "Did you meet anyone during your 6-month guarantee program?" This question is misleading, asking users if they have met *anyone* during their subscription period rather than if they have met their "special someone." The terminology change from the service agreement to the eligibility question tricks users into answering in a way that excludes them from being eligible for the Guarantee Extension.
- Roach Motel: Brignull describes the "roach motel" as when "the design makes it very easy for you to get into a certain situation, but then makes it hard for you to get out of it (e.g. a subscription)."[1] Match.com attracts users to their subscription services with Guarantee Extensions and emails from "interested" users. But, canceling a subscription proves lengthy, including the provision of a password and survey answers. The cancellation process can often exceed *six* page clicks. Users routinely get billed after they believed they canceled their subscription.
- Hidden Information: Match.com frequently hides or fails to disclose relevant information to users. For example, the survey form during the cancellation flow is actually optional: it is possible for users to click through the flow and cancel without providing answers. However, this fact is *never stated* during the cancellation flow, nor are users asked if they want to complete a survey. Moreover, statements such as "Tell us more." and "Before you go, help us make Match.com better" lead users to believe the survey is compulsory. Other important eligibility requirements are embedded in unnumbered paragraphs in the Guarantee Program rules, including the presence of a progress page.

[1] Brignull, *supra.*

- Forced Continuity: This "negative option renewal" prolongs users' service with Match.com after their initial subscription: subscribers are charged automatically for a subsequent term unless they explicitly cancel the subscription. Subsequently, "this pattern takes advantage of users' failure to check up on service expiration dates, either for a free trial or for a limited-time use of a paid service, by **assuming** upon service expiration that the user either wants to continue the paid service…and charges the user."[2]

Total number of steps to cancel
2016:

1. Select settings
2. Select Change/Cancel membership from account settings page
3. Enter password on password screen
4. Page 1 of cancellation flow: Click "Cancel Subscription" from subscription menu options
5.

Notes from Wroblewski, Web Form Design: Filling in the Blanks

"To keep people focused on completing a form, you also should consider which Web site elements help illuminate a clear path to completion and which elements distract from it."
"Removing interface elements not directly related to completing a form helps keep people on task and removes paths to abandonment." (Ch. 3)

Best practices (Ch. 3):
Make sure that you illuminate a clear path to completion through a form by using clear scan lines and effective visual pacing that comfortably takes people from start to finish.
For mission-critical forms like check-out or registration, remove distractions and any links or content that may lead to form abandonment.

Ch. 10: Unnecessary Inputs

- Discussion of removing questions that can apply to the survey questions

"Any question you ask people requires them to parse it, formulate a response, and then input their answer in the affordance you have provided on the form. Being vigilant about every question you ask allows you to remove questions that are not absolutely necessary, or can be asked at a better time or place, or can be inferred automatically."

Research on password friction:

https://uxdesign.cc/15-rules-of-user-sign-in-experience-ae9011d04ee3

[2] Gray, *supra* at 6.

"Unless your site holds sensitive information, allow persistent logins. This is especially true for ecommerce sites. Persistent logins allow the user to experience the site and the actions they've taken. You are a UX criminal if you auto-logout users after a certain time. Sessions may expire, but let the users actions (like items added to cart), remain. You can restrict access to personal information with a password prompt, outside of session expiry. Amazon does this beautifully by keeping you partially logged in and asking for authentication only when you need to access personal information." - useful for making the argument that the "personal info" they are restricting access to is inconsistent both internally compared to other parts of the account, as well as comparatively across other companies

https://www.beyondidentity.com/blog/password-resets-and-the-consumer-journey

- "Half of respondents were likely to leave a site if required to sign-in with a password."
- Includes online dating in their summary of sites where users have to reset passwords

https://www.mcclatchydc.com/news/nation-world/national/article156635539.html
Forgot your password? You have too many and stores are losing business over it By Tim Johnson tjohnson@mcclatchydc.com Updated June 16, 2017 7:03 PM

- Baymard says it sees an 18.75 percent abandonment rate due to reset email issues.

https://baymard.com/research/checkout-usability

- During testing, receiving the **password-reset email** was an especially problematic step, with many participants becoming frustrated and abandoning their task when password emails were delayed or caught in spam filters, or when they had separate issues with signing in to their email account in the first place.
- Indeed, our testing revealed that, in practice, restrictive-password requirements carry serious checkout UX — and, thus, conversion rate — consequences. Some sites during our testing, saw up to 18% abandonment rates among all their existing account users, as these users tried to sign into their existing accounts but couldn't due to forgotten passwords (and ended up abandoning during the password reset process).

- https://www.spiceworks.com/marketing/ecommerce/news/58-consumers-abandon-shopping-carts-due-to-log-in-frustrations-survey-finds/
    - The survey that polled 1,000 consumers in the U.S. revealed that 58% of consumers have abandoned carts and stopped their purchases due to difficulty signing in. Consumers canceled these transactions because they could not remember their password or were being forced to create a new account and password to make the purchase.
    - Note that I can't find the details on this survey, so cite with caution.

Friction, interaction cost, cognitive load:
https://www.nngroup.com/articles/pain-points/

https://www.nngroup.com/articles/interaction-cost-definition/
https://www.nngroup.com/articles/minimize-cognitive-load/

From mathur's paper 'what makes a dark pattern dark':
https://arxiv.org/pdf/2101.04843.pdf

2.3.1 Asymmetric. Asymmetric dark patterns impose unequal burdens on the choices available to the user. The choices that benefit the service are feature prominently while the options that benefit the user are typically tucked away behind several clicks or are obscured from view by varying the style and position of the choice. Asymmetric dark patterns are particularly common in consent interfaces. For instance, the Trick Questions dark pattern can impose a cognitive burden on choices that withhold consent by using confusing language and double negatives. The Confirmshaming dark pattern can use emotion to burden choices, associating guilt with certain choices and not others. Privacy Zuckering burdens choices with user interface friction, hiding privacy-respecting settings behind obscure menus.

The concern about the cognitive burden of an interface has not been cited an explicit normative concern in the dark patterns literature. If the concern is raised, it is in the context of specific dark patterns such as Nagging, Hard to Cancel, and Hidden Legalese Stipulations (and other information overloading dark patterns). As one example, in the context of cookie consent dialogs, Soe et al. [47] argue that cookie consent dialogs without a negative option to deny consent "also introduce additional cognitive burden on the user." Figure 2 shows once instance of a cookie consent dialog with a hard to exercise deny option

The regulatory objectives perspective on dark patterns is more instrumental then normative—the most forceful normative arguments for implementing regulatory objectives are typically the normative arguments for those objectives, rather than compliance with regulation in the abstract. This perspective does not inherently advance a normative argument about why we should care financial losses, privacy harms, or cognitive burdens, beyond noting whether the law directs us to care about those values. This perspective does come with a significant advantage, though: fashioning regulation into measurable metrics for empirical research is usually much easier than adapting normative principles to research.

Evaluation. The regulatory objectives perspective on dark patterns is more instrumental then normative—the most forceful normative arguments for implementing regulatory objectives are typically the normative arguments for those objectives, rather than compliance with regulation in the abstract. This perspective does not inherently advance a normative argument about why we should care financial losses, privacy harms, or cognitive burdens, beyond noting whether the law directs us to care about those values. This perspective does come with a significant advantage, though: fashioning regulation

into measurable metrics for empirical research is usually much easier than adapting normative principles to research.

Caro's paper:
https://www.gmfus.org/sites/default/files/Sinders%2520-%2520Design%2520and%2520Information%2520Policy%2520Goals.pdf

- Don Norman's 1988 Design of Everyday Things helped popularize the term "User-Centered Design."3 His six design principles have since become foundational in the product design space.4 They are:
    - Visibility, referring to how apparent functions are. The more visible functions within a product, the more likely it is for a user to be able to figure out what to do next.
    - Feedback, creating information about an action and what was accomplished.
    - Constraints, referring to how to restrict or select what kinds of interactions a user can do at any particular moment within a product.
    - Mapping, referring to the relationship between users, controls, and the effects of controls in the world. Almost all products have a relationship between controls and effects, be it a light switch, an e-commerce platform, a car, or a flashlight.
    - Consistency, referring to designing interfaces or design choices that have similar operations, interactions, and elements for specific tasks. For example, consistency can be a back button and a forward button placed in the same place throughout a digital experience.
    - Affordance, referring to the attributes of products and how those attributes guide or allow users to know how to use the object. A computer mouse invites touching with buttons but is also constrained to fit into one's hand.5
- Burying choices within multiple steps. Some labels responsive to regulatory requirements are designed to secure consent in ad tracking but are hard to find. For example, some website labels designed to comply with the European Union's General Data Protection Regulation (GDPR) hide "reject" buttons underneath multiple steps.22
- Confusion within sign-up and unsubscribe features. This can be seen in sign-up flows on websites where the user intends to sign up for one subscription but is tricked into signing up for multiple subscriptions and/or products.23
- Design can subvert or thwart policy intentions. If we look to Norman's principles for guidance in building a product, then design should clearly represent how a product functions, with user feedback, clear constraints as to what a product can do, and consistency across interfaces. Dark patterns serve as opposite examples of such principles at work, causing confusion and inconsistency in interfaces, while often not accurately presenting what a product or design is capable of. The

unsatisfactory implementation history of recent regulations offers examples of how dark patterns can subvert policy

Could we use 'ease of use' to help prove that the survey didn't read as optional?
https://www.interaction-design.org/literature/topics/ease-of-use

Apple iOS/Design Guidelines for canceling an account:
https://developer.apple.com/design/human-interface-guidelines/patterns/managing-accounts

If you help people create an account within your app or game, you must also help them delete it, not just deactivate it. In addition to following the guidelines below, be sure to understand and comply with your region's legal requirements related to account deletion and the right to be forgotten.

IMPORTANT

If legal requirements compel your app to maintain accounts or information — such as digital health records — or to follow a specific account-deletion process, clearly describe the situation so people can understand the information or accounts you must maintain and the process you must follow.

Provide a clear way to initiate account deletion within your app or game. If people can't perform account deletion within your app, you must provide a direct link to the webpage on which people can do so. Make the link easy to discover — for example, don't bury it in your Privacy Policy or Terms of Service pages.

DEVELOPER NOTE

If people used Sign in with Apple to create an account within your app, you revoke the associated tokens when they delete their account. See Revoke tokens.

Provide a consistent account-deletion experience whether people perform it within your app or game or on the website. For example, avoid making one version of the deletion flow longer or more complicated than the other. Consider letting people schedule account deletion to occur in the future. People can appreciate the opportunity to use their remaining services or wait until their

subscription auto-renews before deleting their account. If you offer a way to schedule account deletion, offer an option for immediate deletion as well.
Tell people when account deletion will complete, and notify them when it's finished. Because it can sometimes take a while to fully delete an account, it's essential to keep people informed about the status of the deletion process so they know what to expect.
If you support in-app purchases, help people understand how billing and cancellation work when they delete their account. For example, you might need to help people understand the following scenarios:

- Billing for an auto-renewable subscription continues through Apple until people cancel the subscription, regardless of whether they delete their account.
- After they delete their account, people need to cancel their subscription or request a refund.

In addition to helping people understand these scenarios, provide information that describes how to cancel subscriptions and manage purchases. For guidance, see Helping people manage their subscriptions and Providing help with In-App Purchases.

Survey Best Practices
https://www.surveymonkey.co.uk/mp/survey-guidelines/
This suggests emailing ppl info about the survey and the survey:
https://help.surveymonkey.com/en/surveymonkey/policy/data-collection-privacy/

All unsubscribe survey stuff Caro could find

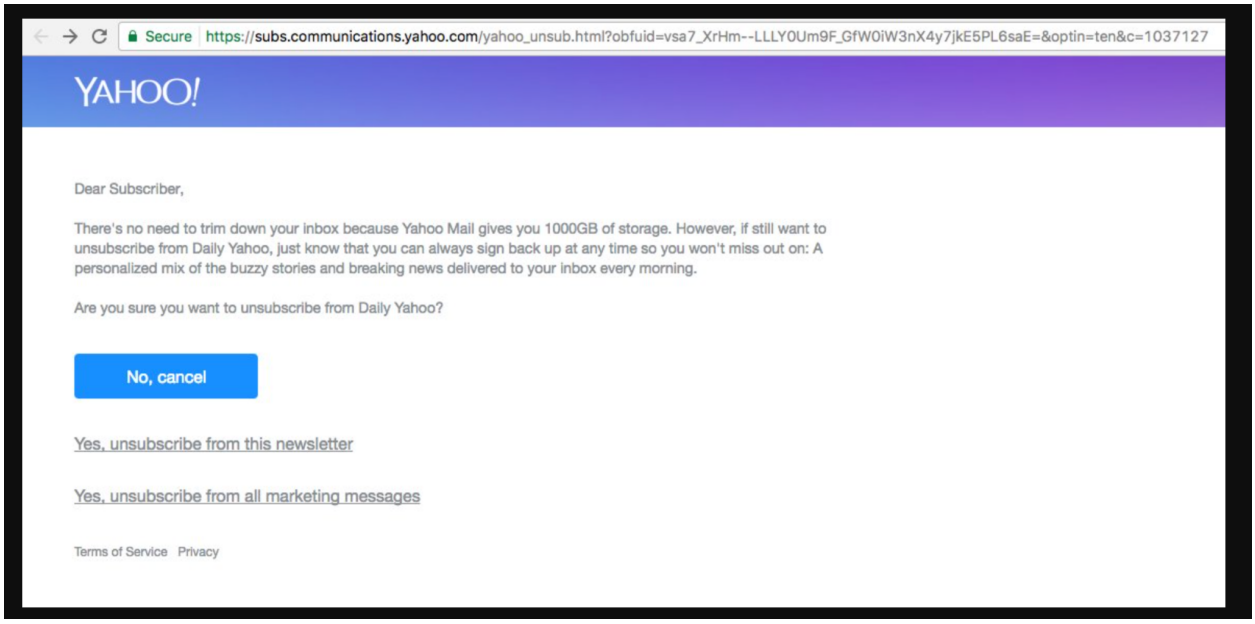
- https://www.campaignmonitor.com/blog/email-marketing/find-out-reasons-for-unsubscribing-with-a-quick-exit-survey/
  - This is for newsletters, fyi (their suggestions)
  - This suggests an optional survey
  - "The quality of your unsubscribe email page can make a difference in increasing subscriber retention rates. The takeaway is not to think of an unsubscribe email message as an unsubscribe exit strategy, but rather a method to improve your content for future customers."
  - All of their email unsubscribe survey examples are ONE interstitial and one question! Just one!

- https://ux.stackexchange.com/questions/17054/should-survey-for-canceling-subscription-be-before-or-after-the-subscription-is
    - One answer suggests a single set of radio dial questions with the un-subscribe (so one question)
        - "If a user goes to the trouble of unsubscribing from something, they *really want to*, and are having some sort of negative experience that they want to alleviate. To retain as much goodwill as possible with the customer, **let them do what they want** -- cancel. To gather a bit of data before they leave, such as the answer to "Why?" then present them with a single set of radio buttons with predetermined answers, and one optional textarea to say more. The act of submitting the form both saves the data point and allows them to complete the action they want to perform. If you really need/want to ask more than one question, then in my experience the most successful response rates are when the questions are presented after the action. Again, this allows the user to complete what it is that *they* wanted to do; a very clear response on the resulting page that their action is complete, and then a clear (brief) appeal to provide feedback, is more likely to produce results than an up-front action blocker."
    - While this is from 2019, a follow up example shows Facebook's delete account option which is just one question
- https://community.hubspot.com/t5/Tips-Tricks-Best-Practices/Implement-unsubscribe-survey-when-someone-unsubscribes/td-p/681735
    - This Q&A with hubspot suggests one question as well as why unsubscribe
- https://www.mailerlite.com/blog/unsubscribe-survey-know-why-your-readers-leave
    - Also suggests using a single question for the unsubscribe survey
- https://mailchimp.com/en-gb/help/edit-or-remove-the-unsubscribe-reason-survey/
    - Mailchimp also shows just 1 question
- https://www.termsfeed.com/blog/unsubscribe-best-practices/
    - This says to let users unsubscribe quickly (but they mean actually unsubscribe users in under 10 days- this isnt about the survey)
    - But the survey example they show from Nordstroms shows the survey being called optional
    - And they suggestion making an optional, short survey
- https://medium.com/@the_manifest/9-tips-for-compelling-email-unsubscribe-pages-e2e8cae01c8f
    - This says don't require the survey, make it optional (but the example they use doesn't show that it's optional)
    - "If your reader is at work or otherwise short on time, requiring them to read through a long list of choices in a survey/poll can chase them away. If you want to collect information on why people unsubscribe from your list, try adding a simple and polite survey on your unsubscribe confirmation page. But, don't force readers to answer questions in order to unsubscribe."

- https://www.litmus.com/blog/the-dos-and-donts-of-unsubscribes/
    - This also says "**Don't**: Make people fill out a survey *before* they've unsubscribed"

Examples of other experts calling or highlighting the visual dark patterns in the Match.com user flow as dark patterns:

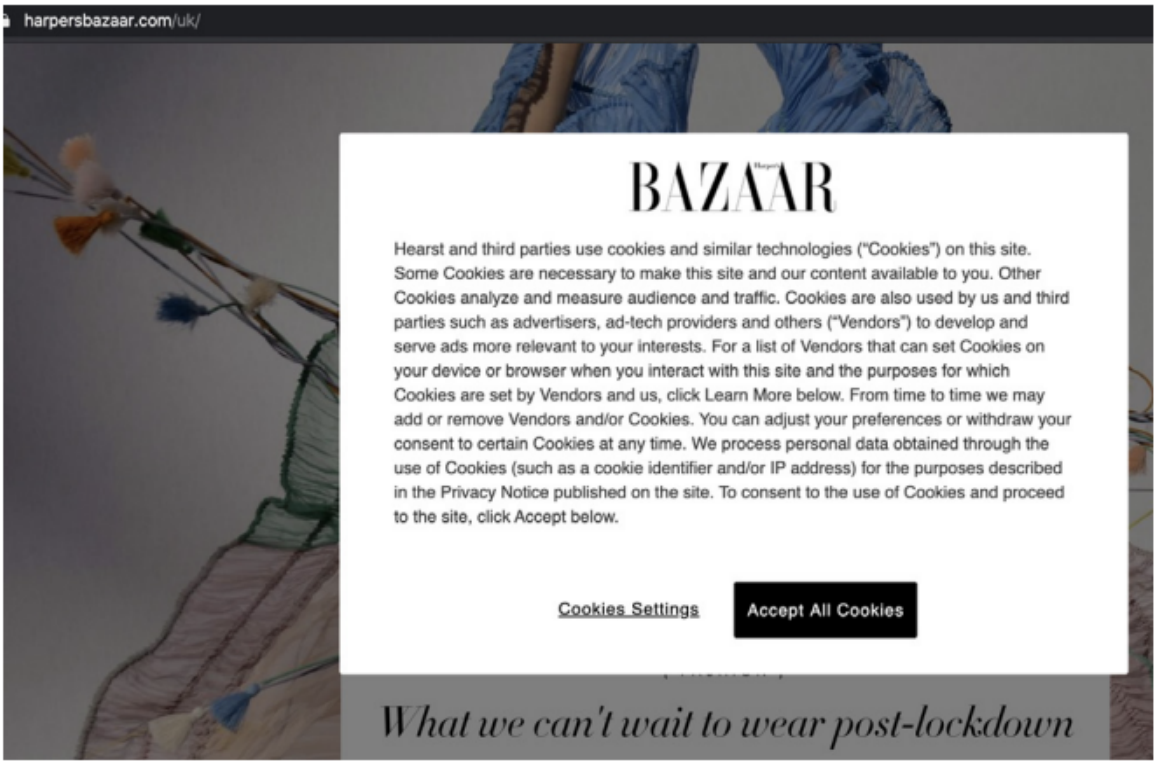
- Yahoo Unsubscribe:



https://darkpatterns.uxp2.com/pattern/yahoo-confusing-unsubscribe/

- Unclear terms of subscription- Adobe- https://darkpatternstipline.org/sightings/unclear-terms-of-subscription/

- Harper's Bazaar cookie banner example:



2.2. Harper's Bazaar's GDPR interstitial, May 2, 2020.

https://www.gmfus.org/sites/default/files/Sinders%2520-%2520Design%2520and%2520Information%2520Policy%2520Goals.pdf

- Is this relevant?
  - https://dl.acm.org/doi/pdf/10.1145/3411764.3445779
  - 2.1.1 Design choices that impact user behavior. In 2019, Utz et al. [96] conducted a field study on more than 80,000 German participants. Using a shopping website, they measured how the design of consent banners influence the behaviour of people acceptance or denial of consent. They found that small UI design decisions (such as changing the position of the notice from top to bottom of the screen) substantially impacts whether and how people interact with cookie consent notices. One of their experiments indicated that dark patterns strategies such as interface interference (highlighting "Accept" button in a binary choice with "Decline"), and pre-selected choices for different uses of cookies has a strong impact on whether the users accept the third-party cookies. In their 2020 study, Nouwens et al. [76] performed a study on the impact of various design choices relating to consent notices, user interface nudges and the level of granularity of options. They scraped the design and text of the five most popular CMPs on top 10,000 websites in the UK, looking for the presence of three features: 1) if the consent was given in an explicit or implicit form; 2) whether the ease of acceptance was the same as rejection—by checking whether accept is the same

widget (on the same hierarchy) as reject; and 3) if the banner contained pre-ticked boxes, considered as noncompliant under the GDPR [44, Recital 32]. In their results, they found less than 12% of the websites they analyzed to be compliant with EU law. In their second experiment, they ran a user study on 40 participants, looking at the effect of 8 specific design on users' consent choices. They recorded an increase of 22 percentage points in given consent when the "Reject all" button was removed from the first page, and "hidden" at least two clicks away from this first page. Finally, they found a decrease of 8 to 20 percentage points when the control options are placed on the first page

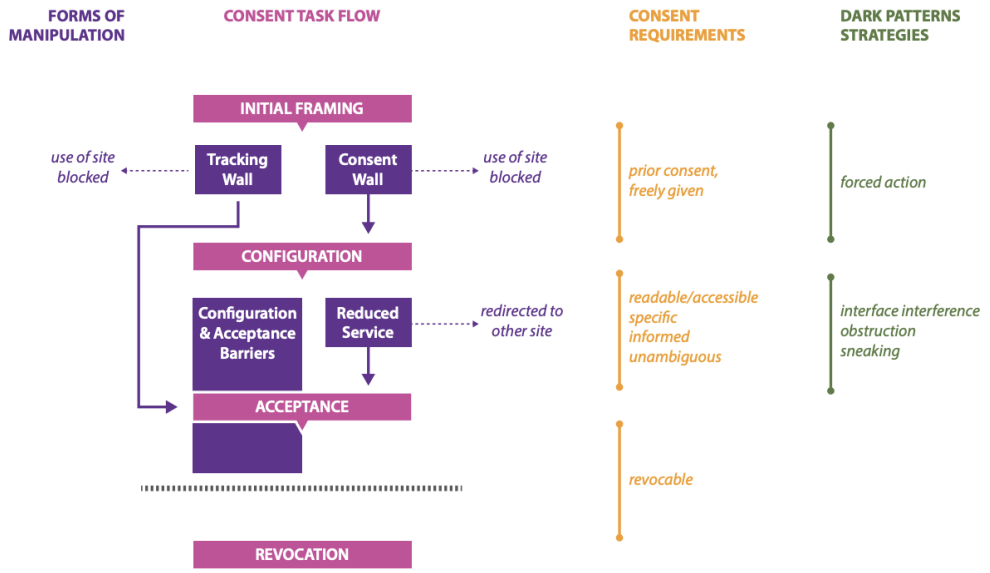


Figure 7: Flowchart describing the forms of manipulation we observed in our dataset in relation to the consent task flow, legal consent requirements, and dark patterns strategies.

Online dating dark patterns (from academic literature and online resources):

- This online website about dark patterns has a comment about Match.com from 2013
  - https://90percentofeverything.com/2013/07/23/the-slippery-slope/index.html
- This is all about matching expectations but nothing on dating
  - https://www.system-concepts.com/insights/persuasive-design-vs-dark-patterns/
- This mentions an FTC and match.com case as an example of dark patterns…"Despite the overall similarity of the distributions in Figure 1, it does give us our first hints of differences across the modalities. First, the app modality produced the longest tail, with apps like Match Dating and Wish containing 18 and 17 dark patterns respectively.

Second, the service in our corpus with the fewest dark patterns overall (USPS Mobile) only contained one pattern in the app modality and none in the two browser modalities, which contributes to the browser modalities starting at 0."
    - https://www.ftc.gov/system/files/ftc_gov/pdf/PrivacyCon-2022-Gunawan-Pradeep-Choffnes-Hartzog-Wilson-A-Comparative-Study-of-Dark-Patterns-Across-Mobile-and-Web-Modalities.pdf
- And from the jackson sun mentioning Match.com
    - https://eu.jacksonsun.com/story/news/2022/04/15/ftc-cracks-down-dark-patterns/7324985001/
- From ACM on match.com
    - https://cacm.acm.org/magazines/2020/9/246937-dark-patterns/abstract
- Mention the adobe.com example here as similar to match.com?
    - https://pacscenter.stanford.edu/wp-content/uploads/2021/07/I-Obscura-Zine.pdf
- Match.com's dark pattern mentioned as a joke here in 2016
    - https://medium.com/@thelonelyrobot/dark-patterns-9748f2b08a95
- First two examples are what Match.com are doing
    - https://www.makeuseof.com/tag/what-are-dark-patterns/

Slightly out of scope but helpful framing stuff [

It is crucial to differentiate between friction that supports mindfulness in users and friction that hinders or holds users hostage within a flow. Friction designed to increase user mindfulness appears in the form of interventions, such as keeping a food diary or photographing meals during one's fitness journey [ ]; challenges or sites of skill enhancement for video game players [ ]; and user-controlled goal; and user-set boundaries for work communications [ ]. These pauses are carefully placed in their interactions to promote users' growth, skill attainment, and quality of life. In these instances, users are provided the tools to breach pauses and control the pace of interaction, aligning them closer to their goals or values. On the other hand, friction that deters users from clearly defined goals (e.g. making a purchase, revoking subscription) leaves the users disempowered, especially when they have little control over when these obstacles are introduced. It is this deterrent friction that is of concern in Match.com's cancellation flow.

- Bait and Switch: Nonsubscribers receive emails indicating a Match.com user has expressed interest in him or her ('bait'). Upon buying a subscription to Match.com new subscribers learn the account that reached out to them was fraudulent or 'unavailable' for viewing ('switch').
- Trick Questions: The Guarantee Program Subscription indicates that users are eligible for a Guarantee Extension (free six-month subscription) if they have not met their "special someone." In obtaining the extension, users are asked, "Did you meet anyone during your 6-month guarantee program?" This question is misleading, asking users if they have met *anyone* during their subscription period rather than if they have met their "special someone." The terminology change from the service agreement to the eligibility question tricks users into answering in a way that excludes them from being eligible for the Guarantee Extension.

]

UNITED STATES DISTRICT COURT
FOR THE DISTRICT OF TEXAS
DALLAS DIVISION
---oOo---

FEDERAL TRADE COMMISSION,
Plaintiff,
vs. No. 3:19-cv-02281-K
MATCH GROUP, INC., a corporation, MATH GROUP, LLC, formerly MATCH.COM, LLC, a Limited Liability Company,
Defendants.
_____/

DEPOSITION OF
JENNIFER KING, PH.D.
CONFIDENTIAL

THURSDAY, JULY 27, 2023

REPORTED BY: HOLLY THUMAN, CSR No. 6834, RMR, CRR
JOB NUMBER 6028094

--o0o--
Videotaped deposition of JENNIFER KING, PH.D., taken by the Defendant, at SIDLEY AUSTIN LLP, 555 California Street, San Francisco California 94104, commencing at 9:02 A.M., on THURSDAY, JULY 27, 2023, before me, HOLLY THUMAN, CSR, RMR, CRR.
--o0o--
APPEARANCES
FOR THE PLAINTIFF:
U.S. FEDERAL TRADE COMMISSION
1999 Bryan Street, Suite 2150
Dallas, Texas 75201
By: M. HASAN AIJAZ, Attorney at Law
MAijaz@ftc.gov
FOR DEFENDANT:
SIDLEY AUSTIN LLP
1999 Avenue of the Stars, 17th Floor
Los Angeles, California 90067
By: CHAD S. HUMMEL, Attorney at Law
CHummel@sidley.com
SIDLEY AUSTIN
2021 McKinney Avenue, Suite 2000
Dallas, Texas 75201
By: CHELSEA A. PRIEST, Attorney at Law
CPriest@sidley.com
ALSO PRESENT:
JEANETTE TECKMAN, In-house counsel, Match.com
SAMUEL KITCHENS, Match.com (Remote)
BRANDON WARD, Precocity (Remote)

I N D E X
INDEX OF EXAMINATIONS

--o0o--
EXHIBITS MARKED FOR IDENTIFICATION

(Cont'd)

(Exhibits, cont'd)

--o0o--
INSTRUCTIONS TO WITNESS/REQUESTS TO MARK TRANSCRIPT

	PAGE	LINE
Instruction not to answer	42	12
Instruction not to answer	66	21

--o0o--

THURSDAY, JULY 27, 2023
9:02 A.M.
--o0o--
JENNIFER KING, PH.D.,
called as a witness, having been first duly sworn, was examined and testified as follows:
---oOo---
EXAMINATION BY MR. HUMMEL
(Deposition Exhibits 1 and 2 were marked for identification.)
BY MR. HUMMEL:
Q. Good morning.
A. Good morning.
Q. My name is Chad Hummel. I represent Match, the defendants in this case.
You understand you're under oath?
A. I do.
Q. And you are providing expert testimony in this case?
A. I am.
Q. And you're being paid by the Federal Trade Commission for that testimony. Correct?
A. I am.
Q. Okay. And you formed some opinions, have you,

hypothetically, and complains about the cancellation flow but it can be verified that they never attempted to enter the cancellation flow, do you consider their complaint valid?
A. Definitely.
MR. AIJAZ: Objection. Calls for speculation.
BY MR. HUMMEL:
Q. Definitely?
A. Yes.
Q. How would -- if they've never even experienced the flow?
A. Because some of those people were calling in because they could not figure out how to enter or even accomplish the cancellation flow.
Q. For those people that had comments about the flow unrelated to their ability to find it but they had never demonstrably entered the flow, do you consider the complaint valid?
MR. AIJAZ: Objection. Vague.
THE WITNESS: Yes. I do consider that to be valid.
BY MR. HUMMEL:
Q. Did you consider the motivation of people calling at all in assessing your complaints?
MR. AIJAZ: Objection. Vague.

THE WITNESS: I mean, what do you mean by "motivation," exactly?
BY MR. HUMMEL:
Q. They were calling to get a refund --
MR. AIJAZ: Objection. Foundation.
BY MR. HUMMEL:
Q. -- as opposed to actually having a complaint about the flow, a bona fide complaint about the flow?
A. So are you suggesting that people are calling and lying?
Q. Yes.
MR. AIJAZ: Objection. Foundation.
BY MR. HUMMEL:
Q. Did you consider that?
A. No. I did not.
Q. Are you an expert in consumer perception?
A. Consumer perception of what?
Q. Consumer perception of things that appear to them on a website.
A. Yes, although that's not how I would describe it, I guess.
Q. Are you a user experience researcher?
A. That's not my title, but I would consider that one of my skills.
Q. Do you have an expertise in user experience

research?
A. Yes, I do.
Q. Do you have a degree in user experience research?
A. I have a Ph.D. in information management and systems with a concentration in humanity-computer interaction, law and policy, and social computing.
Q. What was your thesis?
A. My thesis was on -- my primary area of expertise is in privacy in HCI, and my thesis was on -- it was entitled "Privacy" -- oh, gosh, I forgot the title of my thesis. It's been a while.
Is it in my CV? Let me double-check. It's been five years. As a matter of fact -- sorry -- "Privacy and Social Exchange Theory."
Q. What is that?
What's the Readers Digest version of your thesis title?
What were you exploring?
A. Well, in my CV, I tell you. I used a social relational framework to explore consumer motivations for disclosing personal information to companies.
Q. So your thesis wasn't in website usability. Correct?
A. No. It was not specifically in website

usability issues, but I have published in that area.
Q. Are you an expert web designer?
A. No. I am not a web designer.
Q. Are you an expert app designer?
A. No. By "app" -- well, let me -- what do you mean by "app designer"?
Q. Do you design apps that can be purchased on the --
A. Technical design? Visual design?
(The reporter requested that people not speak at once.)
THE WITNESS: I'm sorry. Technical? Visual? All of the above? I'm not sure where you're going.
BY MR. HUMMEL:
Q. All of the above.
Do you consider yourself an expert in app design?
A. I do not design apps, but I understand a lot about the technical capacity of mobile apps.
Q. Are you an expert graphic designer?
A. No. I do not practice graphic design.
Q. Have you ever designed a website for a paying client?
MR. AIJAZ: Objection. Vague.
THE WITNESS: I have designed a website for an

I don't remember. And I'm just -- from this screenshot, I cannot tell whether both of those options lead to the same place.
Q. Would you agree that you don't get confirmation until after you take a survey?
A. This case, yes. And I -- I noted that on my -- on my report that it was a non-optional survey at this stage.
Q. Right --
A. What appears to be a non-optional survey.
Q. Could you look, please, at page 31 of -- well, just to be clear, on the Facebook and the Coffee Meets Bagel cancellation flows that you examined and opined about in your expert report.
These were examples of good flows that you cited. Correct?
A. No. I mean, they weren't necessarily being held up as models. They were being held up as other dating sites.
I mean, my point was that Mr. Ward's report didn't look at any other dating site other than eHarmony.
Q. Right. He looked at cancellation flows.
A. Yes, across different sites.
Q. And across different industries?

A. Yes, but that he had only one -- one service in the list from a dating site, and that was eHarmony.
Q. Right. Okay. But for the ten that he listed on page 31 of his report, did you go through the cancellation flows personally on each one of those and do a heuristic analysis on each one?
A. I looked at his screenshots; but, no, I didn't conduct a heuristic analysis on each one.
Q. Why?
A. It seemed out of the scope, given that it -- you know, A, the whole point of this was to look at Match.com.
Q. Okay.
A. But, B, I was -- I was concerned that looking at other commercial sites may not be equivalent to looking at other dating sites.
Q. With respect to the complaints that you reviewed for your rebuttal report, did you do anything to confirm that the complaints were true?
A. To confirm that the complaints were true -- in the sense of that they were not fraudulent?
Q. Yes.
A. No. I'm not sure how I would have done that.
I assume that if Match is providing complaints that they have reviewed, that -- that they were not,

let's say, created by spam bots, for example.
Q. But you looked at complaints made that are housed in the FTC Sentinel database. Correct?
A. A small subset. Yes.
Q. What is the primary goal, from the company's perspective, of a cancellation flow?
MR. AIJAZ: Objection. Calls for speculation. Outside the scope.
THE WITNESS: Right. I mean, you could probably argue it's to decrease churn, to use the language you used earlier.
BY MR. HUMMEL:
Q. Well, on the Match.com business model, Match.com exists to allow consumers to meet each other, date, and find love. Right?
That's the point of the site. Correct?
A. I haven't looked at their mission statement, but presumably.
Q. Right. So isn't part of their business model to have users have a cancellation experience that's simple?
In other words, you find love. You don't want to be on Match.com anymore. You want to be able to cancel it, and you want that cancellation process to be simple, easy, so consumers don't have a bad experience

in case the love doesn't work out and they want to come back to Match.com.
Isn't that logical?
MR. AIJAZ: Objection. Vague and foundation.
THE WITNESS: Well, certainly, if you review -- I mean, in the documents I reviewed, there are employees that were trying to make that precise point that, you know, the -- the problems that they spotted in the cancellation flows should be fixed in part because the company was spending a lot of money on excessive customer contact from people who are unhappy and that, yes, because you want to make those people happy if they do resubscribe.
I mean, to the extent -- I agree with those employees that were making that point.
But, of course, they were arguing that because they had found problems with the cancellation flow.
BY MR. HUMMEL:
Q. Well, none of those employees said that the cancellation flow that they were working with at the time was not simple; they just said it could be improved. Right?
MR. AIJAZ: Objection. Characterization of evidence. Foundation.
THE WITNESS: Yeah, I'm not sure I would agree

--o0o--

I declare under penalty of perjury that the foregoing is true and correct. Subscribed at _____, California, this ____ day of _____ 2023.

_____

JENNIFER KING, PH.D.

M. Hasan Aijaz
maijaz@ftc.gov

August 10, 2023

RE: Federal Trade Commision v. Match Group, Inc., Et Al.
7/27/2023, Dr. Jennifer King (#6028094)

The above-referenced transcript is available for review.

Within the applicable timeframe, the witness should read the testimony to verify its accuracy. If there are any changes, the witness should note those with the reason, on the attached Errata Sheet.

The witness should sign the Acknowledgment of Deponent and Errata and return to the deposing attorney. Copies should be sent to all counsel, and to Veritext at errata-tx@veritext.com.

Return completed errata within 30 days from receipt of testimony.

If the witness fails to do so within the time allotted, the transcript may be used as if signed.

Yours,
Veritext Legal Solutions

CERTIFICATE OF REPORTER

I, HOLLY THUMAN, a Certified Shorthand Reporter, hereby certify that the witness in the foregoing deposition was by me duly sworn to tell the truth, the whole truth, and nothing but the truth in the within-entitled cause; that said deposition was taken down in shorthand by me, a disinterested person, at the time and place therein stated; and that the testimony of the said witness was thereafter reduced to typewriting, by computer, under my direction and supervision;

That before completion of the deposition, review of the transcript [X] was [ ] was not requested/offered. If requested, any changes made by the deponent (and provided to the reporter) during the period allowed are appended hereto.

I further certify that I am not of counsel or attorney for either or any of the parties to the said deposition, nor in any way interested in the event of this cause, and that I am not related to any of the parties thereto.

HOLLY THUMAN, CSR No. 6834

Federal Trade Commision v. Match Group, Inc., Et Al.
Dr. Jennifer King (#6028094)

E R R A T A S H E E T

PAGE_____ LINE_____ CHANGE_____
_____
REASON_____
PAGE_____ LINE_____ CHANGE_____
_____
REASON_____
PAGE_____ LINE_____ CHANGE_____
_____
REASON_____
PAGE_____ LINE_____ CHANGE_____
_____
REASON_____
PAGE_____ LINE_____ CHANGE_____
_____
REASON_____
PAGE_____ LINE_____ CHANGE_____
_____
REASON_____

_____ _____
Dr. Jennifer King Date

Michael A. Piazza (SBN 235881)
piazzam@gtlaw.com
Wayne R. Gross (SBN 138828)
grossw@gtlaw.com
Alan A. Greenberg (SBN 150827)
greenbergal@gtlaw.com
GREENBERG TRAURIG, LLP
3161 Michelson Drive, Suite 1000
Irvine, California 92612
Telephone: (949) 732-6500
Facsimile: (949) 732-6501

Attorneys for Defendant
Charles Gugliuzza

UNITED STATES DISTRICT COURT

CENTRAL DISTRICT OF CALIFORNIA

SOUTHERN DIVISION

FEDERAL TRADE COMMISSION,

Plaintiff,

vs.

COMMERCE PLANET, INC., a corporation, and MICHAEL HILL, CHARLES GUGLIUZZA, and AARON GRAVITZ, individually and as officers of COMMERCE PLANET, INC.,

Defendants.

CASE NO. SA CV 09-01324 (CJC) (RNBx)

DEFENDANT CHARLES GUGLIUZZA'S MOTION *IN LIMINE* FOR ORDER EXCLUDING EXPERT TESTIMONY OF JENNIFER KING PURSUANT TO *DAUBERT v. MERRELL DOW PHARMACEUTICALS, INC.*; MEMORANDUM OF POINTS AND AUTHORITIES IN SUPPORT THEREOF

Assigned to: Hon. Cormac J. Carney

Date: May 16, 2011
Time: 3:30 pm
Ctrm: 9B

Trial Date: May 24, 2011

TO PLAINTIFF AND ITS ATTORNEYS OF RECORD:

PLEASE TAKE NOTICE that on Monday, May 16, 2011, at 3:30 p.m., or as soon thereafter as the matter may be heard in the above-entitled Court, located at 411 West Fourth Street, Santa Ana, California 92701, Defendant Charles Gugliuzza will move the Court for a pre-trial order excluding all expert testimony of Jennifer King ("King") from being submitted at trial on the grounds that her proposed testimony is improper legal opinion, fails to assist the trier of fact, and lacks reliability under Federal Rules of Evidence 701 and 702 pursuant to *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579, 113 S. Ct. 2786, 125 L.Ed. 2d 469 (1993).

This motion is made following the Local Rule 7-3 required conference of counsel which took place on April 22, 2011 and this Court's Order granting a shortened briefing schedule. *See* Docket No. 122. The motion will be based on this Notice of Motion and Motion, the Memorandum of Points and Authorities in support thereof, Declaration of Wayne R. Gross, all other pleadings and papers on record with this Court, and upon such arguments and items as may be presented to the Court at the hearing of this matter.

DATED: April 29, 2011

GREENBERG TRAURIG, LLP

By: */s/ Wayne R. Gross*
Michael A. Piazza, Esq.
Wayne R. Gross, Esq.
Alan A. Greenberg, Esq.
Attorneys for Defendant,
Charles Gugliuzza

TABLE OF CONTENTS

TABLE OF AUTHORITIES

MEMORANDUM OF POINTS AND AUTHORITIES

I. INTRODUCTION

In this motion *in limine*, Defendant Charles Gugliuzza ("Gugliuzza") seeks an order excluding Plaintiff Federal Trade Commission ("FTC") and its attorneys from making any reference to, or offering any evidence relating to, the opinions of FTC witness Jennifer King ("King"). Defendant anticipates that the FTC will seek to introduce her as a "social technologist" -- a term she coined -- who will seek to opine that "most" consumers of Online Supplier would not know that they were entering a negative option.

As will be demonstrated, King possesses no expertise in negative option disclosures. Moreover, even if she did possess the requisite expertise, her opinions are entirely unreliable and therefore should be excluded on that basis as well.

II. RELEVANT FACTS

A. COMPLAINT AND MISLEADING ATTACHMENT

In November 2009, the FTC sued Charles Gugliuzza claiming that he, as president of a holding company, knowingly and/or recklessly caused a subsidiary to disclose insufficiently on its website the terms of a negative option membership program. There is no dispute that the website did disclose the membership terms and its recurring fees. The FTC's lawsuit, however, contended that the disclosures were inadequate and that Gugliuzza knowingly or recklessly participated in a fraud on consumers. Specifically, the complaint alleged in pertinent part as follows:

> Although [the website] contained language to indicate that the transaction "involves a negative option" and that consumers "may be liable for payment of future goods and services. . . for $59.95 per month . . . ." in most standard screen configurations that language appeared below the bottom of screen. *See* Exhibit 2, page 7. Because the disclosure language appeared below the bottom of the screen, it was possible for consumers to complete the entire transaction without ever having seen it. As a

> consequence, consumers may not, and in a substantial number of instances did not, understand that they had been enrolled in a negative option continuity plan, that is, an agreement according to which consumers will be billed monthly for services, whether they used those services or not, unless they affirmatively sought to discontinue the plan.

(Ex. B[1], Complaint ¶ 18).

Attached to the complaint as Exhibit 2 was a purported depiction of the website page referenced in Paragraph 18. (*Id.*) However, the single page of the website was depicted in seven different hard copy pages, with the first page of the exhibit depicting no more than approximately half of the top of the page, and each successive page depicting lower portions, until finally, on page seven, there appeared a depiction of the negative option disclosure. (*Id.*) The obvious purpose of displaying the website page in seven different hard copy pages was to make it appear consistent with the FTC's allegation that "the disclosure language appeared below the bottom of the screen."

B. DEPOSITION OF CHIEF TECHNOLOGY OFFICER

After the complaint and during the discovery phase of the case, the FTC turned over a declaration signed by Ethan Brooks, the former Chief Technology Officer of Commerce Planet. (Ex. C, Declaration of Ethan Brooks.) The declaration, which discussed various aspects of website at issue, made no mention of the placement of the negative option disclosure referenced in the complaint. (*Id.*)

On January 6, 2011, Mr. Brooks was deposed by counsel for Gugliuzza and was questioned about how the declaration was created. (Ex. D, Deposition Transcript of Ethan Brooks, 62:2-63:25.) Mr. Brooks explained that, in October 2010, FTC counsel requested a meeting with Mr. Brooks at a hotel. (*Id.* at 25:14-26:11.) At the meeting, FTC counsel, utilizing a laptop, showed Mr. Brooks a video of what the FTC claimed

[1] All Exhibits referenced herein are appended to the accompanying Declaration of Wayne R. Gross in Support of Defendant Charles Gugliuzza's Motions *in Limine* for Order Excluding Expert Testimony of Molly Petullo and Jennifer King ("Gross. Decl.").

was a screen capture of the website sign-up page. (*Id.* at 54:2-55:23.) The video made it appear that one would have to scroll the page to view the terms and conditions. (*Id.*) After the video was played, FTC counsel referenced the fact that the terms and conditions were beneath the fold of the page. (*Id.* at 56:9-57:5.) In response, Mr. Brooks informed FTC counsel that, in the video, the screen had been set at too low of a resolution and that, had the screen resolution been set correctly, the terms and conditions would be viewable without scrolling. (*Id.* at 57:6-58:7.)

During the deposition, Mr. Brooks was shown the FTC complaint, including the exhibit that contained seven different hard copy pages, successively depicting the sign-up page from top to bottom. (*Id.* at 153:15-157:20.) Mr. Brooks testified that the exhibit misrepresented the viewable portion of sign-up page as it would be viewable on a computer screen. (*Id.*)

Mr. Brooks additionally testified that, after the meeting at the hotel, FTC counsel sent Mr. Brooks a draft declaration for his signature. (*Id.* at 91:8-11.) The declaration made no mention of either the video or the explanation by Mr. Brooks that, contrary the FTC's allegation, the terms and conditions were above the fold. (*Id.* at 91:8-99:21.) Mr. Brooks, after making minor revisions to what was contained in the declaration, signed it. (*Id.*)

C. RETENTION OF JENNIFER KING AS A TESTIFYING EXPERT

In September 2010, the FTC retained Ms. King as a consulting expert in the case against Mr. Gugliuzza. At the time, Ms. King, a self-described "social technologist" -- did not possess any expertise pertaining to negative option disclosures or the placement of such disclosures on a website, and did not even know what negative options were prior to being retained. (Ex. E, Deposition transcript of Jennifer King, 16:20-23, 27:7-28:4.) Nevertheless, the FTC, in January 2011 -- the same month in which Brooks was deposed -- changed Ms. King status from a consulting expert to a testifying expert. (*Id.* at 10:19-21.) In her capacity as a testifying expert, Ms. King rendered an opinion that "most" consumers who signed up for the membership would not know they had entered a

negative option. (*Id.* at 72:2-20, 203:15-25.) She rendered this opinion without interviewing a single customer or conducting any empirical analysis. (*Id.* at 75:12-14; 106:10-19.)

III. ARGUMENT

A. LEGAL STANDARD GOVERNING EXPERT TESTIMONY

The opinions of Jennifer King are subject to the same standards of reliability that govern the expert opinions of strictly scientific experts retained for the purposes of litigation. *See Kumho Tire Co., Ltd. v. Carmichael*, 526 U.S. 137, 151 (1999) (holding that *Daubert* applies even when an expert's opinion relies on skill or experience-based observation). Under Federal Rules of Evidence Rule 702 and *Daubert*, the Court is to ensure that any and all expert testimony or evidence admitted is not only relevant, but reliable. Rule 702 of the Federal Rules of Evidence states:

> If scientific, technical or other *specialized knowledge* will assist the trier of fact to understand the evidence or determine a fact in issue, a witness qualified as an expert by knowledge, skill, experience, training, or education, may testify thereto in the form of an opinion or otherwise, if (1) the testimony is based upon sufficient facts or data, (2) the testimony is the product of reliable principles and methods, and (3) the witness has applied the principles and methods reliably to the facts of the case.

Id. (emphasis added).

Rule 702 imposes upon courts the obligation to act as gatekeepers, by ensuring all expert testimony, whether scientific, technical, or any other specialized knowledge, is both reliable and relevant. *See United States v. Freeman*, 498 F.3d 893, 901 (9th Cir. 2007). Courts engage in a two-part analysis in determining the admissibility of an expert opinion. See *Ralston v. Smith & Nephew Richards, Inc.*, 275 F.3d 965, 969 (10th Cir. 2001). First, the Court must determine whether the expert is qualified by knowledge, skill, experience, training, or education to render an opinion. *Id.* Second, the Court must determine whether the expert's opinions are sufficiently reliable. *Id.*

B. JENNIFER KING IS AT MOST AN EXPERT ON PRIVACY ISSUES AND THEREFORE ANY EXPERTISE SHE POSSESSES HAS NOTHING TO DO WITH THIS CASE

The Supreme Court has held that any expert testimony offered during trial is required to exhibit a clear connection between the expert's "knowledge" and the "pertinent inquiry" at issue. *See Daubert v. Merrell Dow Pharm., Inc.*, 509 U.S. 579, 591-92 (1993). Gugliuzza does not dispute that an appropriate expert on consumer behavior and marketing can opine in this case on issues pertaining to whether consumers were likely to see and understand the negative option disclosures at issue. *See Federal Trade Commission v. Amy Travel Services, Inc.*, 875 F.2d 564, 574 (7th Cir. 1989) (court upheld magistrate judge's finding that testimony on how consumers would react to sales material should be given by an expert in consumer psychology or consumer behavior). Indeed, the defense has retained Ken Eisner, an expert on these topics, for that very purpose.[2] However, it is abundantly clear that Jennifer King -- a self-described "social technologist" -- lacks the requisite expertise.

On April 18, 2011, King was deposed. During the deposition, King admitted that she had never been retained as an expert previously, had no expertise in FTC compliance or negative option disclosures, and would not feel comfortable holding herself out to private industry as an expert in negative options disclosures. (Ex. E at 11:8-12:4, 16:14-23, 187:2-188:1.) Indeed, the following colloquy established her scant knowledge of the concept:

> Q: All right. Have you ever been hired by a company to provide the company with advice on how to -- to do a negative option disclosure?

[2] *See* Ex. F (expert report of Kenneth J. Eisner dated March 7, 2011) and Ex. G (Rebuttal of Jennifer King's Expert Report by Kenneth J. Eisner dated March 21, 2011).

A: No. Actually, before I was introduced to this case I wasn't even particularly -- I had not heard the term -- the term of art "negative option disclosure."

Q: Okay. So prior to September of 2010 you had never heard of the term "negative option disclosure"?

A: Not in a way -- not in a meaningful way not I didn't know what it was off the top of my head. I thought I knew what it was. I'm certainly familiar with the types, but as a term of practice, I had not heard negative option disclosure in my work.

Q: Okay. And so -- and I know this may be hard because now, obviously, you do. But prior to September of 2010 what did you think negative option was, if you remember?

A: Well, I thought it was a very odd term.

Q: And I know -- again, I want you to separate?

A: Yeah.

Q: What you now know from what you knew prior to September of 2010. If you can't do it, tell me, but if you can, please answer the question.

A: I had a vague sense of what it was, and I would say the closest kind of example I had was the Columbia record and tape club, which I was a member of when I was a child. So I knew it had something to do with that kind of realm and practice, but I didn't know much more than that.

(*Id.* at 27:23-29:5.)

Thus, King had no more than a vague notion of what negative options were in September 2010, when the FTC chose to retain her as a consulting "expert" for a case pertaining to that very topic. To make matters worse, King, even by the time of her deposition in April 2011, admitted that she is still not an expert in the placement of negative option disclosures and would not feel comfortable holding herself out as one to the private sector. (*Id.* at 11:8-12:4, 16:14-23, 187:2-188:1.)

So what expertise does she possess? Her CV explains it as follows: "My primary research is an empirical inquiry (utilizing both qualitative and quantitative methods) of how technology impacts information privacy for individuals and institutions." (Ex. H, curriculum vitae of Jennifer King, at 1). Her CV proceeds to list eleven publications, each of which relates to privacy issues. (*Id.*) The first one cited illustrates this point: "How Different are Young Adults from Older Adults When it Comes to Information Privacy Attitudes and Policies?" (*Id.*) The CV also lists seven "Invited Talks & Panels" which, like her publications, all relate to privacy issues. (*Id.*) During her deposition, she admitted that she has neither written nor spoken on negative option disclosures. (Ex. E at 27:4-22.) This makes sense, of course, given that she did not even know what negative option disclosures were prior to being retained by the FTC.

Her lack of expertise -- or even knowledge -- of negative option disclosures is particularly problematic in a negative option disclosure case given that neither the FTC nor private industry has offered specific rules or best practices on how negative options should be disclosed. For example, Ms. King, during her deposition, stated that, to her knowledge, she knew of no rule that would preclude a company from placing a negative option disclosure at the bottom of a webpage.

> Q: Okay. And as far as you know, the FTC has never passed a rule that precludes a company from placing a negative option disclosure at the bottom of a landing page, correct?
>
> A. Have they given specific guidance not to do that; is that what you're asking?
>
> Q. (Nods head.)
>
> A. No, not to my knowledge.

(*Id.* at 117:13-16.)

Later in the deposition, Ms. King further admitted that private industry similarly has not offered guidance, such as best practices, on how negative options should be disclosed.

Q: We've already discussed that the FTC has not created specific rules regarding how a negative disclosure is to be done, correct?

A: Yes.

Q: Okay. Is it fair to say that industry -- private industry has not adopted a specific best practice with respect to how a negative disclosure should be done?

A: To the best of my knowledge, they have not.

(*Id.* at 125:17-126:1.)

The absence of rules or specific guidance on the placement of negative option disclosures underscores the importance of allowing only experts with appropriate expertise testify on whether such disclosures are understood by consumers. That King does not possess such expertise was reflected in her responses to the following questions:

Q: Given that there are no specific rules from the FTC regarding negative disclosures and private industry has not adopted, to your knowledge, best practices regarding negative disclosures, how would a company in 2006 and 2007 know that a particular negative option disclosure violated FTC rules?

A: Well, I think you would have to first start with the circa 2000 guide, which you printed here, Exhibit 377. And the, working with the guidelines they give, I think you can default to basic human-computer interaction principles at that point and work through -- work through it from that perspective, which is how I approached it in my report.

Q: So what you're saying is that -- and we'll go through your report in detail, but that a company, in order to know whether they were violating the FTC laws would have to apply certain academic principles?

A: Not necessarily academic ones. Certainly some of the research I cite is based in academia, the books you just brought out, for example, though occupy a space between academic research and practitioner research, and

this is a field where there are a lot of practitioners who contribute to the general body of knowledge.

Q: Is it fair to say that how to make a particular negative option disclosure is a gray area?

Mr. Rose [FTC Counsel]: Objection. Vague.

A: How to make a particular negative option disclosure that it's a gray area? It's gray inasmuch as there's no book off the bookshelf one can go buy that says here's how to do a negative option disclosure.

Q: Which is why you didn't cite any books that provided clear indication as to how a negative option disclosure should be done, correct?

A: Certainly, yes.

(*Id.* at 126:14-128:4.)

As reflected by this colloquy, the FTC chose an "expert" who utterly lacks the appropriate expertise to opine on a topic that she herself admitted is a gray area and where there is a lack of written resources -- either from government or private industry -- that would enable her to quickly learn about the topic after being retained for this case. She was and still is at most a privacy expert, a topic that has nothing to do with this matter. Her testimony should be disqualified on this basis alone.

C. KING'S OPINIONS ARE NOT RELIABLE

Even if King did possess the requisite expertise, her opinions are utterly unreliable. Prior to rendering her "opinion" that "most" consumers would not understand the negative option disclosure, King did not interview a single consumer and testified that the FTC had not either. Nor did King, unlike the defense expert, conduct or rely upon any empirical analysis. Rather, King's "opinion" was based on the assumption that the disclosure was not viewable to consumers absent scrolling and that "most" consumers would not scroll, even though she admitted that she would herself. (*Id.* at 72:2-20, 269:24-270:8, 88:21-89:1.) King formulated this assumption from reviewing an article that purported to set forth the frequency with which various screen resolutions were

utilized by computer users during the relevant timeframe. (*Id.* at 193:19-195:5.) Her assumption, however, ignored the fact that more factors than screen resolution determine whether one would need to scroll to see a particular website page. Her deposition revealed this point as follows:

> Q: And it varies from -- potentially from computer to computer as to how much of these pages are viewable to a reader; is that correct?
>
> A: It varies based on screen resolution.
>
> Q: Does it also vary based upon the size of the computer screen?
>
> A: Yes, it can also vary based upon that.
>
> Q: All right. And does it also vary based upon the use of tool bars?
>
> A. It can. It can.
>
> Q: Okay
>
> A. It can make the viewable portion smaller.
>
> Q: Right. And so, sitting here today, you don't know with precision how much of each landing page could be viewed by Online Supplier consumers in 2006 and 2007, correct?
>
> A. With precision? No.
>
> Q: Right. And you haven't interviewed any consumers of Online Supplier, correct?
>
> A: I have not.

(*Id.* at 74:20-75:14.)

To make matters worse, King disregarded key evidence reflecting that, contrary to her assumption, the terms and conditions were above the fold. During her deposition, Ms. King admitted that, contrary to her "expert" opinion, Commerce Planet's former Chief Technology Officer had testified that the terms and conditions were in fact viewable to Online Supplier consumers without scrolling. (*Id.* at 191:3-14). Even more significantly, Ms. King, in rendering her opinion that the terms and conditions were below the fold, failed to consider indisputable documentation that reveals that Gugliuzza

knew that the terms and conditions were above the fold. As set forth in his declaration in support of motions for summary judgment, Gugliuzza explained his knowledge and the corroborating documentation as follows:

> I am aware that in February 2007, Paul Huff (Commerce Planet's In-House Counsel) oversaw changes to the Online Supplier webpages. Although I was relying on Mr. Huff to ensure legal compliance, I also saw these pages myself and the terms and conditions were clearly set forth. I am now aware in this lawsuit that the FTC is alleging that the terms and conditions were "below the fold" and that user would need to scroll down to see them. While I do not think that the need to scroll makes a webpage deceptive in any way, I remember seeing the terms and conditions "above the fold" and assumed that is how everyone saw it. Attached as Exhibit "BB" is an email I sent on April 5, 2007, showing a screen shot I made of the Online Supplier credit card entry page. It clearly shows that on my computer, which I never set to any special resolution or display settings, the terms and conditions regarding the negative option and monthly charges is set forth in plain view "above the fold." Prior to the FTC investigation, no one ever told me that the terms and conditions were "below the fold" or any less conspicuous than it appeared on my computer screen.

(*See* Docket No. 108, ¶ 16.)

In sum, King's so-called expert opinions are no more reliable than her expertise in negative option disclosures. Both lack the substance mandated under Rule 702 and should be excluded.

IV. CONCLUSION

Based upon the foregoing, Gugliuzza respectfully requests an Order prohibiting the FTC from presenting any testimony or evidence regarding Jennifer King, and that the

FTC, its attorneys, and all other persons attending or participating at the trial on its behalf be admonished not to refer to these matters during the trial.

DATED: April 29, 2011

GREENBERG TRAURIG LLP

By: /s/ Michael A. Piazza
Michael A. Piazza, Esq.
Wayne R. Gross, Esq.
Alan A. Greenberg, Esq.
Attorneys for Defendant

UNITED STATES DISTRICT COURT

CENTRAL DISTRICT OF CALIFORNIA

HONORABLE CORMAC J. CARNEY, JUDGE PRESIDING

CERTIFIED TRANSCRIPT

- - - - - - -

FEDERAL TRADE COMMISSION, PLAINTIFF(S), VS. COMMERCE PLANET, INC. ET AL, DEFENDANT(S).	)	NO. SACV 09-1324-CJC DAY FIVE

REPORTER'S TRANSCRIPT OF PROCEEDINGS

COURT TRIAL

SANTA ANA, CALIFORNIA

TUESDAY, FEBRUARY 07, 2012

MARIA BEESLEY-DELLANEVE, RPR, CSR 9132
OFFICIAL FEDERAL REPORTER
RONALD REAGAN FEDERAL BUILDING, ROOM 1-053
411 WEST 4TH STREET
SANTA ANA, CALIFORNIA 92701
(714) 564-9259

2012-02-07

APPEARANCES OF COUNSEL:

FOR THE PLAINTIFF(S): FEDERAL TRADE COMMISSION
BY: DAVID NEWMAN, ESQ.
AND ERIC EDMONDSON,
EVAN ROSE, ESQ.
KERRY O'BRIEN, AAL
901 MARKET STREET, SUITE 570
SAN FRANCISCO, CALIFORNIA 94103
(415) 848-5100

FOR THE PLAINTIFF(S): FEDERAL TRADE COMMISSION
BY: RAYMOND MCKOWN, ESQ.
10877 WILSHIRE BLVD., SUITE 700
LOS ANGELES, CALIFORNIA 90024
(310) 824-4343

FOR THE DEFENDANT(S): GREENBERG TRAURIG, LLP
BY: WAYNE GROSS, ESQ.
AND MICHAEL PIAZZA, ESQ.
ALAN GREENBERG, ESQ.
3161 MICHELSON DRIVE, SUITE 1100
IRVINE, CALIFORNIA 92612
(949) 732-6500

I N D E X

PLAINTIFF'S WITNESS	DIRECT	CROSS	REDIRECT	RECROSS	VOIR DIRE
JENNIFER KING		4	81	97	
MICHAEL HILL	110	179			

EXHIBITS

PLAINTIFF'S EXHIBIT	DESCRIPTION	FOR IDENTIFICATION	IN EVIDENCE
6			126
13			148
14			148
27			166
28			167
177			142
259, 261			154
1035			132
1050			160
1182			176
1246			127
1247			153
1289			175
1331			150

09:54 A DIDN'T DO WHAT HERE?
09:54 Q THAT TYPE A STUDY.
09:54 A CONDUCT A STUDY TRYING TO MEASURE THAT PARTICULAR ASPECT, NO, I DID NOT.
09:54 Q YOU ALSO DIDN'T CONDUCT A PHYSIOLOGICAL STUDY. AND BY THAT I MEAN EYE TRACKING. YOU DIDN'T ATTEMPT TO PUT UP A WEB PAGE TO A USER AND TRACK THEIR EYES TO SEE WHERE THEY LOOK ON THE PAGE, WHICH IS SOMETHING THAT CAN BE DONE; CORRECT?
09:54 A IF YOU HAVE A GOOD SENSE OF WHO THE CUSTOMERS ARE, SO THAT YOU CAN RECRUIT A REPRESENTATIVE GROUPING OF THEM, YES, YOU COULD.
09:55 Q YOU HAD ACCESS TO CUSTOMER COMPLAINTS; RIGHT?
09:55 A I HAD ACCESS TO AT LEAST SOME. I DON'T KNOW IF THEY WERE ALL.
09:55 Q THOSE WOULD BE POTENTIAL ACTUAL USERS; RIGHT?
09:55 A POTENTIALLY.
09:55 Q AND YOU MADE NO EFFORT TO CONTACT THEM; RIGHT?
09:55 A NO, I DID NOT.
09:55 Q YOU WOULD AGREE THAT, AT LEAST AS A PART OF A TEST GROUP, THAT MIGHT BE A LIKELY GROUP TO TEST PHYSIOLOGICALLY; RIGHT?
09:55 A IT WOULD PROBABLY BE BIASED. I WILL RATHER HAVE A NON-COMPLAINING GROUP OF CUSTOMERS, JUST A GROUP OF POTENTIAL CUSTOMERS WHO FIT WHAT THE COMPANY THOUGHT WERE THEIR TARGET AUDIENCE WITHOUT NECESSARILY SELECTING FROM A GROUP THAT'S COMPLAINING.
09:55 Q I WAS TRYING TO ANTICIPATE THAT. I SAID THEY MIGHT BE PART

09:55 OF A TEST GROUP, NOT THE ENTIRE TEST GROUP.
09:55 A THAT WOULDN'T BE FAIR TO MAKE THEM THE ENTIRE TEST GROUP.
09:56 Q IN THE FIELD OF HCI, THERE IS ACTUALLY NO PROHIBITION -- I WANT TO GET BACK TO THE WEB PAGES THEMSELVES.
09:56 THERE IS NO PROHIBITION ON SCROLLING, IS THERE?
09:56 A NO. THERE'S DEFINITE CIRCUMSTANCES IN WHICH SCROLLING IS APPROPRIATE. ANY NEWSPAPER WEBSITE PROVIDES A GOOD EXAMPLE. THEY COULDN'T GET ALL OF THE TEXT YOU WOULD NEED TO READ ON SCREEN. YOU WOULD ANTICIPATE NEEDING TO SCROLL IN THAT SENSE BECAUSE YOU HAVE AN ENTIRE ARTICLE WEAVING ITS WAY DOWN THE PAGE.
09:56 Q LET'S TALK ABOUT HOW YOU DID YOUR USER INSPECTION --
09:56 A USABILITY INSPECTION?
09:56 Q -- USABILITY INSPECTION OF THE WEB PAGES.
09:56 WHAT BROWSER DID YOU USE?
09:56 A I USED SEVERAL. I USED FIREFOX, I LOOKED AT THEM IN SAFARI, THESE WERE BOTH ON MAC, AND I LOOKED AT THEM BOTH IN FIREFOX AND INTERNET EXPLORER IN WINDOWS.
09:57 Q SAFARI DIDN'T EXIST IN '06 AND '07; RIGHT?
09:57 A NO. THAT WAS JUST TO MAKE SURE I WASN'T HAVING ANY PARTICULAR RENDERING ISSUES ON THE MAC.
09:57 Q BUT THAT WOULDN'T GIVE YOU --
09:57 A NO, I DIDN'T DO ANY OF MY BASIS FOR ANALYSIS USING SAFARI. I JUST MENTIONED THAT I CHECKED IT OUT.
09:57 Q FIREFOX HAS BEEN OUT FOR A LONG TIME, BUT IT'S NEVER REALLY GAINED MORE THAN MAYBE 5 PERCENT OF THE MARKET SHARE?

-OOO-

CERTIFICATE

I HEREBY CERTIFY THAT PURSUANT TO SECTION 753, TITLE 28, UNITED STATES CODE, THE FOREGOING IS A TRUE AND CORRECT TRANSCRIPT OF THE STENOGRAPHICALLY REPORTED PROCEEDINGS HELD IN THE ABOVE-ENTITLED MATTER.

DATE: FEBRUARY 7, 2012

/S/ MARIA BEESLEY

Digitally signed by MARIA BEESLEY
DN: cn=MARIA BEESLEY, o, ou, email=amaria1957@yahoo.com, c=US
Date: 2012.02.07 20:25:23 -08'00'

OFFICIAL COURT REPORTER

How to Run an Effective Heuristic Evaluation

Doug Bonderud

Published: June 07, 2023

Heuristic evaluations measure interface usability to pinpoint areas for improvement. These evaluations occur before an application or service is tested by customers.

End-users offer valuable perspectives. However, they may focus on development elements that rely on more critical components.

Instead, experts evaluate interfaces using an established list of criteria during a heuristic evaluation. They then flag interface issues for remediation. One common guide was created by Molich and Nielsen.

In this piece, we'll break down the basics of heuristic evaluations, examine their benefits, then offer best practices to help you deliver actionable insights.

Table of Contents

What is a heuristic evaluation?

A heuristic evaluation is a product development test where experts measure the usability of an interface against an accepted list of principles, known as heuristics. Product developers can conduct heuristic evaluations throughout the development process to ensure the interface of a website or app is ideal for the target consumer.

Heuristic evaluations provide product development teams with an expert assessment of their website's usability. After the inspection, evaluators will give developers and designers a list of potential issues to address.

Product management can then instruct their teams to tweak the interface according to those recommendations. If performed correctly, this process can address over **80% of the usability issues** on your website or app.

Heuristic Evaluation vs. Usability Testing

Both of these tests point out flaws with usability. However, heuristic evaluations differ significantly from **usability testing**. These assessments are conducted differently and identify different types of issues.

Heuristic evaluations are conducted by industry professionals who find flaws based on preset guidelines. The evaluators inspect the interface on their own terms, then provide the development team with a list of suggestions.

Conversely, usability tests observe target consumers while they're using the website or app. These tests give the user a specific task to perform.

Evaluators then observe whether the user can complete the action and how long it took to do so.

During usability testing, participants may be asked to provide some feedback to the developers. However, these insights will be in response to any questions the development team has for the user.

When to Use a Heuristic Evaluation

A heuristic evaluation can be used at any point in the development process.

However, it's most effective when conducted early on in the website or app's design stages.

If possible, heuristic evaluations should even be performed after each **design sprint**. This gives your team useful feedback about your design before users are exposed to it during testing.

Heuristic evaluations also tend to be less expensive to conduct when the interface is in the early stages of development. The more advanced your interface becomes, the more expensive it will cost to redesign it.

By running your heuristic evaluations early and often, you can ensure usability and avoid costly redesigns.

Image source

The Benefits of Conducting a Heuristic Evaluation

There are many usability tests your company can conduct. However, heuristic evaluations provide unique insights that can play a major role in the success of your website or app.

Additionally, they can be much more cost-effective and efficient compared to other testing methods.

This should be enough to sway most product teams. If you're still skeptical about heuristic evaluations, consider these three benefits.

1. Efficiency

Heuristic evaluations, in practice, are a relatively simple process to conduct.

Depending on the product's complexity, they can be completed in as little time as a couple of days.

The experts analyzing the interface often work independently. This allows developers to focus on other projects while the evaluators work.

Once the evaluation is complete, designers can then address the errors found in testing. After corrections are made, they can present another version for evaluators to re-test.

This creates an efficient **feedback loop** that continues throughout the development process.

2. Organization

The feedback from a heuristic evaluation can influence how team prioritizes sprints and projects.

Evaluators provide product management with a list of flaws that are organized by their severity. **Product owners** can then use this information to create and organize their product backlogs.

By using this system for prioritization, product teams are more likely to stay organized and meet their deadlines.

3. Versatility

Heuristic evaluations aren't a one-and-done analysis. Their findings can be used alongside other usability tests to uncover fresh insights.

For example, after you address the feedback from a heuristic evaluation, you can check out your product usage reports to measure the success of your changes.

If you notice areas of lower usage, you can then point out those aspects to evaluators.

Heuristic evaluations then provide product developers with qualitative feedback. This can help explain trends appearing in product usage reports.

Running a Heuristic Evaluation

The specifics of a heuristic evaluation will vary based on the type of service or application you're testing. However, there is a set of seven common steps that can help you run an effective evaluation.

We'll explore each step below.

How to Run a Heuristic Evaluation

1. Determine what you're testing.
2. Clearly define context and goals.
3. Select a team of evaluators.
4. Choose your heuristics.
5. Give evaluators specific instructions.
6. Conduct multiple evaluations.
7. Collect results.

1. Determine what you're testing.

The first step in any heuristic evaluation is to determine exactly what you're testing. This means identifying the interface you're evaluating and the specific usability aspects you want to evaluate.

By pinpointing the target of heuristic evaluation up-front, you can save time and effort down the line.

2. Clearly define context and goals.

The first step in any heuristic evaluation is to determine exactly what you're testing. This means identifying the interface you're evaluating and the specific usability aspects you want to evaluate.

By pinpointing the target of heuristic evaluation up-front, you can save time and effort down the line.

3. Select a team of evaluators.

Next, choose your team of evaluators. Ideally, you're looking for evaluators that have conducted previous assessments in your industry.

While there's no "ideal" number of evaluators, you need at least two to limit the risk of bias. More than 10 evaluators can make it harder to track outcomes.

4. Choose your heuristics.

With your goals and evaluators in place, it's time to choose your heuristics. If you're stuck, you can rely on a common framework, such as Molich and Nielsen.

Evaluators are often asked to rate:

- The simplicity of dialogue in user interfaces.
- The consistency of function and form.
- Efforts taken to provide shortcuts.
- Good error messages.
- Clearly marked exits.

5. Give evaluators specific instructions.

Before turning evaluators loose, provide them with specific testing instructions. This includes training in what will be tested, what rating scales will be used, and how issues will be flagged.

6. Conduct multiple evaluations.

On their first evaluation, teams should freely use the application to identify potential areas for analysis. Their second evaluation can then focus on assessments and issue reporting.

Depending on the complexity of your application, you may want to consider a third evaluation for more in-depth results.

7. Collect results.

Finally, collect evaluators' reports and debrief them in a shared session. This will help pinpoint any common issues and create a plan for remediation.

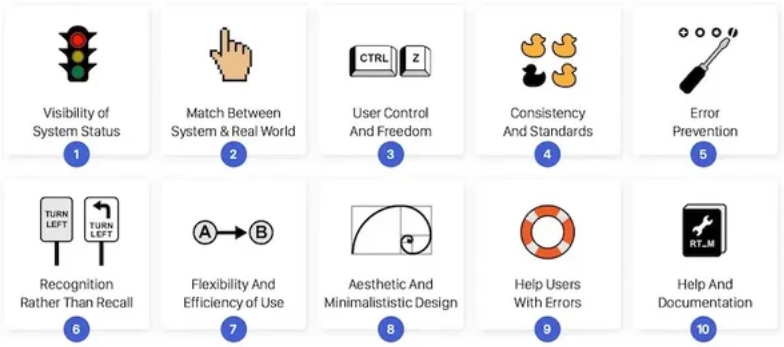
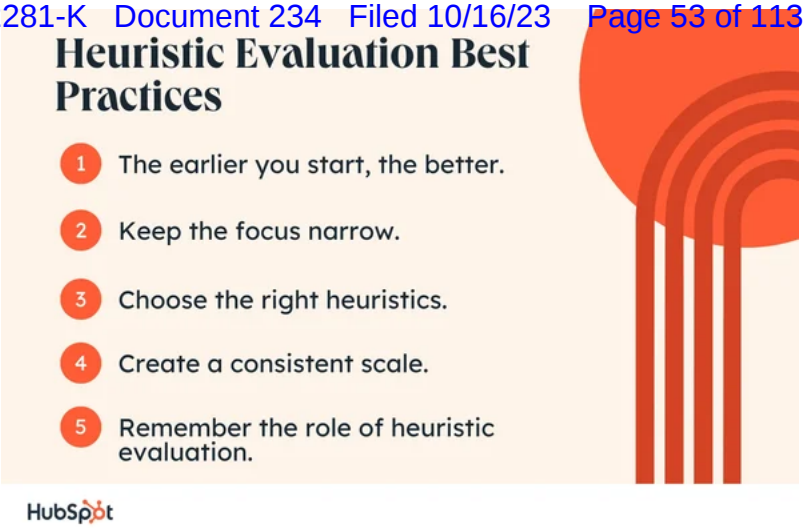


Image source

Heuristic Evaluation Best Practices

Use these five best practices to ensure accuracy and consistency in your heuristic evaluations.



1. The earlier you start, the better.

The earlier you can conduct a heuristic evaluation, the better. Ideally, this means deploying evaluators just after functional prototypes are completed. Because the product is still in its early stages, making changes takes less time and effort.

2. Keep the focus narrow.

Zero in on what you want to test. This might be a set of functions or even a single function. By choosing a narrow approach, you can get more usable data from your testing. You can then make meaningful changes as early as possible.

3. Choose the right heuristics.

Not every heuristic will be right for your evaluation.

For example, one of Molich and Nielsen's heuristics is the provision of clearly marked exits when users click on the wrong link.

If you're evaluating products that have just come out of prototyping, these exits may not be in place. This measurement, therefore, provides no value.

4. Create a consistent scale.

Whether you choose a 1-10, percentage, or Y/N scale, keep it consistent across evaluators and functions. This makes it easier to collect data and prioritize issues.

5. Remember the role of heuristic evaluation.

Last but not least? Remember that heuristic evaluation isn't a replacement for end-user testing.

Heuristic evaluations help identify key interface usability issues. However, this narrow focus means can't provide the same kind of broad-scope insight as other end-user evaluations.

Be sure you pair heuristic evaluations with other, helpful testing.

Fostering Feedback With a Heuristic Evaluation

Heuristic evaluations are the first stop on the road to a complete application or service. Expert evaluators and clear criteria can find widespread usability issues early. This allows companies to make the most of customer assessments.

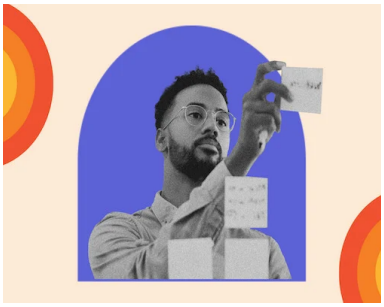
Put simply, if you want to foster consistent, actionable feedback, your first step is a heuristic evaluation.

Editor's note: This post was originally published in March 2019 and has been updated for comprehensiveness.

Topics: User Testing

Don't forget to share this post!

Related Articles



The UX Designer's Guide to Affinity Diagrams

May 25, 2023

Beta Testing: The Ultimate Guide For Product Teams

Apr 27, 2023

Scrum Product Owner: Rol Responsibilities, Explainec

Apr 13, 2023

AN EMPIRICAL STUDY OF USABILITY TESTING: HEURISTIC EVALUATION VS. USER TESTING

Enlie Wang Barrett Caldwell
Industrial Engineering, Purdue University
West Lafayette, Indiana

In this study, two different usability-testing methods (Heuristic Evaluation and User Testing) were selected to test the usability of a pre-release version of software searching for Science, Mathematics and Engineering education materials. Our major goal is to compare Heuristic Evaluation and User Testing in terms of efficiency, effectiveness and cost/benefit analysis. We found that Heuristic Evaluation was more efficient than User Testing in finding usability problems (41 vs. 10), while User Testing was more effective than Heuristic Evaluation in finding major problems (70% vs.12%). In general, Heuristic Evaluation appears to be more economic in finding a wide range of usability problems by incurring a low cost in comparison to User Testing. However, User Testing can provide more insightful data from real users such as user's performance and satisfaction.

INTRODUCTION

The Science Math and Engineering Learning Technologies (SMELT) software prototype is an online educational resource repository designed for teachers, students, and administrators in K-12 schools. Many of these users do not have a reliable Internet connection or enough time and Internet experience to find good educational resources on the web. Thus, SMELT was developed as a solution for these problems. It can help users quickly assess the resources they wanted with minimal search effort. In order to improve the usability and quality of this special educational software, we decided to use two different usability-testing methods to conduct parallel usability testing for the pre-release version of SMELT. Although improving usability of SMELT software is our ultimate goal, comparing the effectiveness and efficiency of different usability testing methods is our major research interest in this study.

A recent literature review in usability testing (Wang, Caldwell and Salvendy, 2002) summarized major usability testing methods (see table 1). Usability testing methods can be classified into four categories (Nielsen 1994a): automatic, usability measures computed by running usability inspection software during the evaluation task; informal, based on rules of thumb and general skill and experience of evaluators; empirical, tested by real users; and formal, using exact models and formulas to calculate usability measurements. After examining the strengths and weaknesses of each, two usability-testing methods (Heuristic Evaluation and User Testing) were chosen as optimal testing methods for testing in this study.

Table 1. Usability methods taxonomy (expended from Nielsen 1994c)

Method Classification	Representative Methods
Automatic	Webtesting software, such as WebArch by *Addwise*, and Web Trends by *NetIQ*
Informal	Heuristic Evaluation (Nielsen 1990) Cognitive Walkthroughs (Lewis, Wharton and Rieman, 1990) Pluralistic Walkthroughs (Bias, 1994) Guidelines, Questionnaires (i.e. QUIS) and Checklists (i.e. Lin, Choong and Salvendy, 1997) Feature Inspection (Bell, 1992)
Empirical	User Testing (response time, error rate and subjective judgment), User Feedback
Formal	GOMS (Card, Moran & Newell 1983, Gong and Kieras 1994) Thinking aloud analysis (Simon 1989, Lewis 1982, Wright 1992)

Heuristics Evaluation

Heuristic Evaluation is done by having usability experts look at an interface and trying to come up with a judgment about what is good and bad about the interface. The most popular heuristics are: *simple and natural dialogue, speak the user's language, minimize user memory load, be consistent, provide feedback, provide clearly marked exits, provide shortcuts, good error messages, prevent errors, help and documentation* (Nielsen and Molich, 1990). This testing technique is intuitive, inexpensive, and quick way of getting a fairly comprehensive usability problem report. Because of these characteristics, it is easy to motivate both management and evaluators to apply this usability testing technique. This method was chosen because the SMELT development team needed quick feedback for redesigning and improving the software before the official release date.

User Testing

Heuristic Evaluation is a quick and inexpensive method, but it is not perfect or without defect. Many studies (Nielsen, 1992, 1993a, 1994a) reported that it tended to find too many false positives (false alarms) and minor problems. A lot of the developer's time and effort may be spent working on these less critical problems. User Testing is the best way to identify what are real problems that will impact user's performance and preference. Because the problems found in User Testing are identified by actual users, there are fewer false positives identified, and all problems found are worth investigating. It also can measure user's performance and satisfaction such as delay tolerance during human-computer interaction (Wang and Caldwell, 2002). However, it takes time to prepare the testing materials and to recruit test subjects. Bailey (1996) did a study to compare the User Testing and Heuristic Evaluation directly by evaluating a series of systems with both methods. He reported that User Testing performed better than Heuristic Evaluation, and each method tended to find different types of usability problems. We also want to examine this claim through our study and compare Heuristic Evaluation and User Testing by evaluating a real software product. Thus, User Testing became the second testing method in this study.

METHODS

Heuristic Evaluation

Participants. The experiment was conducted with 5 graduate students who were, at the time, taking a Usability Testing and Evaluation Techniques seminar at University of Wisconsin-Madison. Five evaluators are sufficient to conduct Heuristic Evaluation according to Nielsen's research (Nielsen 1994 c). Evaluators had no previous experience with the SMELT software before testing. A small gift was given to each evaluator upon the completion of evaluation.

Apparatus. The experiment was run on a Pentium 3 class computer with the SMELT software installed locally on a hard drive. Instruction form and the ten heuristics list were given to the evaluators as guidelines during evaluation. The ten heuristics we used in this study are: (1) visibility of system status, (2) match between system and the real world, (3) user control and free, (4) consistency and standard, (5) error prevention, (6) recognition rather than recall, (7) flexibility and efficiency of use, (8) aesthetic and minimalist design, (9) help users recognize, diagnose, and recover from errors , and (10) help and documentation (Nielsen 1994 b).

Test Procedures. It took the evaluator about 1.5 hours to complete one evaluation session. The evaluation included following steps:

1. Introduction to the SMELT software. The experimenter gave a brief description about the software and explained the functions in the software. Next, the evaluators were asked to familiarize themselves with the software. The whole process took about 15 minutes.
2. The evaluation. Two main parts of the software were evaluated: the "User's Guide" section and the "Search" section. The evaluators were given a set of heuristic principles. They were asked to inspect dialogue elements and compare them with the heuristic principles. Then, they were asked to write the problems found and the heuristic principles that were violated by each problem. Each evaluation lasted about 30 minutes.
3. Aggregate results. After the evaluation was done each evaluator was asked to give general comments about the software, and then the results were aggregated.

Usability Testing

Participants. Ten participants participated in the User Testing. They came from a wide scholastic background including instructors and students at both high school and college levels. Academic disciplines represented Engineering, Food Science and other disciplines. This variety of backgrounds was thought to be more representative of the SMELT user population (see Table 2. Participant background). User's previous

Table 2. Participant Background

Participant Classification	Number	Age (Mean)	Experience (Mean)		
			Computer	Internet	Search Engine
Student	5	21.5 (6.5)	4.2 (2.2)	4.6 (1.1)	4.4 (0.9)
Teacher	5	26.8 (2.8)	5.0 (0.7)	4.8 (1.1)	5.0 (1.4)
Male	5	27.6 (3.1)	4.2 (1.9)	4.4 (0.9)	4.4 (0.9)
Female	5	30.6 (5.3)	5.0 (1.2)	5.0 (1.2)	5.0 (1.2)
Total	10	24.1 (5.5)	4.6 (1.6)	4.7 (1.1)	4.7 (1.2)

computer and software knowledge and experience may effect user's satisfaction and performance toward new software. Three kinds of related experience, computer, internet and search engine experience were measured by seven-point scale ranging form 1(no experience) to 7 (very experienced). The results show that our users have adequate experience in these aspects.

Apparatus. The test was run on an IBM compatible laptop computer (Pentium II, 400 MHz) with the SMELT software installed locally. A video camera recorder was used to record users' activities during testing. A series of forms were developed for this testing process: *Consent Form, General Instructions, Background Survey, User Testing Instructions, Test Task, Formal Tasks Result Recording Sheet and User Satisfaction Survey.*

The user satisfaction survey is composed of two parts. The first part is adapted form of the QUIS 5.0 survey from (Chin, Diehl and Norman, 1988), and the second part is developed by authors to measure users' views of task time, user help, user expectations, memory effort, and search effort.

Test Procedures. A standard procedure was developed that included 9 separate steps: (1) Sign consent form, (2) Complete background survey, (3) Read general instructions, (4) Learn and practice software operation, (5) Read test instructions, (6) Execute test task, (7) Execute four formal simulation tasks (search for online educational resources in a specific area and recommend the best Web site for the users in the simulation scenario), (8)Complete satisfaction survey, (9) Report problems.
In order to limit initial learning effects, users were given enough time to explore the software and the User's Guide before testing. The test was not started until the users indicated that they were familiar with the software. After the formal tasks, the user was to complete user satisfaction survey and report the problems they found during testing. Users were also given access to the SMELT software at this time to aid in recalling usability problems. Each testing session lasted about 1 hour.

RESULTS

Usability Problems and Types

The five evaluators using Heuristic Evaluation found 103 problems in total (before aggregating). On average, each evaluator found 20 problems. After aggregation, 58 unique problems were identified. However, the subjects in User Testing only reported 10 usability problems. In order to study the problem types and severities, we used Nielsen's five-point severity-rating scale (Nielsen, 1994c) to evaluate the problems found by both methods. The scores (0-4) stand for *false alarm, cosmetic problem, minor usability problem, major usability problem,* and *usability catastrophe* respectively. Results showed that among the 58 problems identified by Heuristic Evaluation, there were 5(8.7%) major problems, 18(31%) minor problems, 18(31%) cosmetic problems, and 17(29.3%) false alarms. The ten usability problems found by User Testing were classified into two types: 7(70%) major problems and 3(30%) minor problems; no problem was rated as a false alarm. Table 3 listed some examples found by Heuristic Evaluation and User Testing.

More User Testing Results

The quantitative data collected in this study includes task completion times and user satisfaction (from the surveys). Table 4 summarized task completion times according to tasks and user groups. T-Test showed that no significant difference was found in learning-time or task-time between teachers and students, or between males and females. The average time to complete a search task was 2.24 min (SD= 0.87). These results indicated that SMELT software is easy to learn and use for both teachers and students.

Table 3. Some examples of the problems identified

Problem Type	Heuristic Evaluation	User Testing
Major	"It is difficult to switch to SMELT from other application when multiple applications are running simultaneously because there is no software icon on the task bar "	" Confused by the difference between ***Target*** and ***User***"
Minor	"User's Guide doesn't explain the function of ***Accurate Search***"	"Database is too small"
Cosmetic	"The term ***More Details*** is not explained in User's Guide"	NA
False Alarm	"I prefer green background."	NA

Table 4. Analysis of task completion times

	Learning	Test	Task1	Task2	Task3	Task 4
Students	4.68 (1.92)	2.78 (1.60)	3.52 (1.78)	2.89 (1.64)	1.07 (0.36)	2.20 (0.93)
Teachers	6.21(2.26)	3.95 (0.14)	2.84 (0.57)	2.11 (0.47)	1.18 (0.22)	2.11 (0.48)
Male	5.99(2.35)	4.10 (0.36)	2.74 (0.62)	2.02 (0.50)	1.07 (0.17)	1.98 (0.49)
Female	5.08(1.96)	2.64 (1.40)	3.62 (1.71)	2.98 (1.58)	1.18 (0.41)	2.34 (0.88)
Total	5.54(2.10)	3.37 (1.23)	3.18 (1.30)	2.50 (1.21)	1.13 (0.30)	2.16 (0.70)

Table5. Break down analysis of user satisfaction

Dimensions	Teacher	Student
Part one (9-point scale)		
Overall	6.36 (0.52)	6.88 (1.79)
Learning	7.12 (0.97)	7.84 (1.01)
Terminology	7.32 (0.90)	7.08 (0.33)
Screen	7.50 (1.38)	7.75 (1.06)
System Capabilities	7.20 (1.37)	7.30 (0.84)
Part two (7-point scale)		
Learning time	4.60 (1.14)	5.40 (1.52)
Expectation	6.40 (0.55)	6.60 (0.89)
User's guide	3.60 (1.14)	4.20 (1.10)
Help	4.00 (0.71)	4.20 (0.45)
Quality of results	5.20 (0.84)	4.80 (1.64)
Description	5.20 (1.30)	5.00 (1.00)
Memory effort	4.80 (1.79)	4.80 (1.92)
Search effort	5.80 (0.84)	5.20 (1.30)

The reliability of our usability survey was fairly good (a=0.91, 32 items) and comparable to the original QUIS 5.0 survey (a= 0.94, 27 items). For this study, we divided our survey into two parts: the first part focuses on general criteria such as *Overall impression, Learning, Terminology, Screen and System Capabilities;* while the second part deals with more specific features including *Learning time, User's expectation, User's guide, Help, Quality of search results, Description, Memory effort and Search effort.* In general, users are fairly satisfied with SMELT software. They rated SMELT software 7.18 in the first part (9-point scale) and 4.99 in the second part (7-point scale). However, we also noticed that the overall rating in the first part was relatively lower than other criteria. It was the *User's guide* (3.9 out of 7) and Help (4.1 out of 7) that affected user's overall rating. A break down analysis is provided in the Table 5.

Cost comparisons

Although all the subjects participated in both usability tests were volunteers, we still found that Heuristic Evaluation was cheaper than User Testing in terms of direct cost and time cost. The direct office supply costs including testing materials and gifts for testing users were: $10.54 for Heuristic Evaluation; $47.30 for User testing. It only took experimenter 15.5 hours to conduct the HE including data analysis, but User Testing need 45 hours to complete the whole process.

DISCUSSION

Looking at the results of the cost analysis, Heuristic Evaluation seems to be a more appealing usability technique than User Testing. The Heuristic Evaluation ended up with less expense and less time invested than the User Testing technique. Accordingly, Heuristic Evaluation appears to be more economic in finding a wide range of usability problems by incurring a low cost in comparison to User Testing. Nielsen (1993b) once pointed out user testing was 4.9 times as expensive as the cheapest heuristic method but provided better performance estimates. The cost comparison of this study supports Nielsen's conclusion.

According to Cost Comparison, Heuristic Evaluation was found to be more *efficient* than User Testing, identifying a larger number of problems for a smaller cost in both time and money. However, as graduate students, the evaluators did not demand the salary that an experienced, professional usability expert would be paid. Our results, therefore, may not be particularly accurate in an off-campus setting. We also found some problems while using Heuristic Evaluation to detect usability problems. First, it yields a significant false alarm rate. Second, once a person becomes an expert, he/she will not think and behave like a novice user. Thus, some problems associated with learning are hard to be identified. For example, the test users in User Testing identified two major problems that had not been identified by Heuristic Evaluation. Based on observations made during this study, Heuristic Evaluation is a useful testing method in the earlier stages of software development. It was found that Heuristic Evaluation is a quick method and identifies a wide range of problems. As is widely known, it is easier and cheaper to correct errors in earlier design stages than in the later design stages. We recommended that software designers and developers use Heuristic Evaluation to eliminate as many usability problems as possible in the earlier stages. When a functional prototype of the software is available, User Testing should be considered as major testing method because it can capture real usability problems from the real users. This can help software development team to better meet the needs and expectations of the user population before release of the software.

ACKNOWLEDGMENTS

Portions of the work presented in this paper were supported by grants from the National Institute for Science Education (funded by NSF) and the Wisconsin Space Grant Consortium (funded by NASA), while the authors were at the University of Wisconsin-Madison. The opinions and findings represent the authors' own perspectives, and do not reflect official positions by these or any other agencies.

REFERENCES

Bailey, W. R. (1996). *Human performance engineering: Designing high quality, professional user interfaces for computer products, applications, and systems* (3rd ed.). Upper Saddle River, NJ: Prentice Hall PTR.

Chin, J.P., Diehl, V.A., & Norman, K.L. (1988). Development of an instrument measuring user satisfaction of the Human-Computer Interface. *ACM CHI'88 Proceedings*, 213-218.

Lin, X. H., Choong, Y., & Salvendy, G. (1997). A proposed index of usability: A method for comparing the relative usability of different software systems. *Behaviour & Information Technology,* 16 (4/5), 267-278.

Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. *Proc. ACM CHI'90 Conf.,* 249-256.

Nielsen, J. & Phillips (1993a). Estimating the relative usability of two interfaces: Heuristic, formal, and empirical methods compared. *Proc. ACM CHI'93 Conf.,* 214-221.

Nielsen, J. (1993b). *Usability Engineering*, Boston, MA: Academic Press.

Nielsen, J. (1994a). Usability inspection methods. *Proc. ACM CHI'94,* Boston, Massachusetts USA, April 24-28.

Nielsen, J. (1994b), *Jakob Nielsen's Online Writings on HE.* Retrieved January 16, 2000, from http://www.useit.com/papers/heuristic/.

Nielsen, J. (1994c). Enhancing the explanatory power of usability heuristics. *CHI'94 Con.,* 152-155.

Rubin, J. (1994). *Handbook of Usability Testing: How to plan, design, and conduct effective tests.* New York: John Wiley & Sons, Inc.

Wang, E., & Caldwell, B.S. & Zhang K. (2002). Time delay tolerance in Computer Supported Cooperative Work. *Proceeding of 6th International Scientific Conference on Work With Display Unit,* 199-201.

Wang, E., Caldwell, B.S. & Salvendy, G (2002). Usability Comparison: Similarity and differences between E-commerce and World Wide Web. *Journal of Chinese Industrial Engineering.* Accepted for Publication.

FINDING USABILITY PROBLEMS THROUGH HEURISTIC EVALUATION

Jakob Nielsen
Bellcore
445 South Street
Morristown, NJ 07962-1910
nielsen@bellcore.com

ABSTRACT

Usability specialists were better than non-specialists at performing heuristic evaluation, and "double experts" with specific expertise in the kind of interface being evaluated performed even better. Major usability problems have a higher probability than minor problems of being found in a heuristic evaluation, but more minor problems are found in absolute numbers. Usability heuristics relating to exits and user errors were more difficult to apply than the rest, and additional measures should be taken to find problems relating to these heuristics. Usability problems that relate to missing interface elements that ought to be introduced were more difficult to find by heuristic evaluation in interfaces implemented as paper prototypes but were as easy as other problems to find in running systems.

Keywords: Heuristic evaluation, Interface evaluation, Usability problems, Usability expertise, Discount usability engineering, Telephone-operated interfaces.

INTRODUCTION

Heuristic evaluation [17] is a method for finding usability problems in a user interface design by having a small set of evaluators examine the interface and judge its compliance with recognized usability principles (the "heuristics"). Heuristic evaluation thus falls into the general category of usability inspection methods together with methods like pluralistic usability walkthroughs [1], claims analysis [2][3][10], and cognitive walkthroughs [11][19], with the main difference being that it is less formal than the other methods and intended as a "discount usability engineering" [13][16] method. Independent research has found heuristic evaluation to be extremely cost-efficient [8], confirming its value in circumstances where limited time or budgetary resources are available.

The goal of heuristic evaluation is the finding of usability problems in an existing design (such that they can be fixed). One could thus view it as a "debugging" method for user interfaces. The present article extends previous work on heuristic evaluation [4][12][14][17] by looking more closely at several factors that may influence the probability of finding usability problems. A probabilistic approach is necessary in examining the success of a method that is heuristic and approximate. The factors considered below are the expertise of the evaluators, the severity of the usability problems, the individual heuristics, and the activities needed to identify the problems.

EFFECT OF THE EVALUATORS' USABILITY EXPERTISE

Heuristic evaluation was originally developed as a usability engineering method for evaluators who had some knowledge of usability principles but were not necessarily usability experts as such [17]. Subsequent research has shown the method to be effective also when the evaluators are usability experts [4][8]. Unfortunately, usability experts are sometimes hard and expensive to come by, especially if they also need to have expertise in a particular kind of application.

To investigate the effect of having evaluators with varying levels and kinds of expertise, a study was conducted where the same interface was subjected to heuristic evaluation by three groups of evaluators: "Novice" evaluators with no usability expertise, "regular" usability specialists, and "double" usability specialists who also had experience with the particular kind of interface being evaluated.

A Telephone Operated Interface

A "voice response" system is a computer information system accessed through a touch tone telephone. The user's only input options are the twelve buttons found on a regular telephone (the digits **0–9** and the special characters ***** and **#**). The system's only output is through speech and sometimes sound effects. This interaction mechanism provides literally hundreds of millions of terminals to any computer system and allows it to be accessed from almost anywhere in the world [6][7].

Because of the variety of evaluators employed in the present study, a printed dialogue was evaluated instead of a running system. The evaluators were given a dialogue that had been recorded from a voice response system which will be referred to here as the BankingSystem. Evaluating an interface on the basis of a written specification is actually a reasonable task, and is one of the strengths of the heuristic

evaluation method. It lends itself to such evaluations as well as to evaluations of implemented systems [14].

The BankingSystem is a telephone operated interface to the user's bank accounts. The user's task in the sample dialogue was to transfer $1,000 from the user's savings account to the user's checking account. The dialogue between the BankingSystem (S) and the user (U) in Figure 1 is took place as the user tried to perform this task. This dialogue has actually taken place, the underlying problem being that the user had not authorized the bank to accept transfers over the phone.

The user can be assumed to be provided with printed instructions stating that the system uses the # key to signify the end of the user's input (in the same way as many other systems use an enter key). As long as the user has not hit the # key, it is possible to correct input by pressing ** (the asterisk key used twice). This option is not used in the dialogue in this example, however. The printed instructions were not evaluated as part of the heuristic evaluation.

For the heuristic evaluation, the evaluators were asked to keep in mind those basic technical limitations of the system which were due to the access through a touch tone telephone and not to include criticism of the very use of 12-button input and speech output instead of, say, input though speech recognition or output through graphical dialogues with pull-down menus, etc.

Even the small sample dialogue in Figure 1 contains a rather large number of usability problems as listed below. Readers who are only interested in the general results of the heuristic evaluation and not the individual usability problems can skip ahead to the section titled *Results of the Heuristic Evaluation.* The list of usability problems is ordered according to a rough estimate of the relative seriousness of the problems. Major usability problems are those that have serious potential for confusing users or causing them to use the system erroneously while minor problems may slow down the interaction or inconvenience users unnecessarily. For each problem a brief phrase in italics indicates what well-established usability principle was broken by the dialogue with respect to that problem. These phrases are taken from the list of nine usability heuristics used in earlier papers on heuristic evaluation [12][17] and are discussed in further detail in [16].

Major Problems

Problem 1. The error message in statement 11 appears much too late. It is not reasonable to have the user enter large amounts of information only to learn that it was all in vain. The error message should have appeared immediately after the user's initial choice of the function in statement 2. One evaluator noted that the current form of the dialogue seems to be designed as a front end to a traditional full screen mainframe system where users fill in an entire screen of data before hitting enter and transmitting all of the data to the computer at once. Even if the bank's main computer system with the account database were indeed a traditional transaction-oriented mainframe, one should not let the properties of the backend computer impact the dialogue when a new interaction medium is introduced. In many cases, the best solution would be to prevent the error from occurring in the first place by only listing those functions in the main menu which the user is authorized to use. *(Simple and natural dialogue).*

(First there is a short dialogue in which the user is identified by entering an identification number and a secret access code—this is not part of the evaluation exercise)

(1) S: Enter one for account information, three for transfers between your own accounts, ...

(2) U: **3#** *{the user interrupts the system}*

(3) S: Enter account to transfer from

(4) U: **1234567890#** *{savings account number}*

(5) S: Enter account to transfer to

(6) U: **#** *{an abbreviation for the checking account}*

(7) S: Enter amount in cents

(8) U: **100000#**

(9) S: From account number twelve thirtyfour fiftysix seventyeight ninety to account number primary account, a transfer of one thousand dollars is to be made. Press one to confirm, zero to cancel.

(10) U: **1#**

(11) S: You do not have access to use this function.

Figure 1 *A sample dialogue between a user (U) and the BankingSystem (S). The statement numbers are not part of the dialogue but are used to refer to individual statements in the analysis.*

Problem 2. Users should not be required to convert an amount of money to cents since only a very mathematically inclined person will find it easy to think of $1,000 as 100,000 cents. This problem can be completely avoided by simplifying the system to allow transfer of whole dollar amounts only. Doing so will also speed up the interaction by eliminating two keystrokes. For transfers between the user's own accounts, whole dollars will be completely adequate. For a bill-payment system, it might still be necessary to have some mechanism for specifying cents. In that case, a solution might be to have users enter first the dollars and then be prompted for the cents. Since the system allows the user to interrupt prompts, a transfer of a whole dollar amount could still be achieved very fast if the user entered **1000##** to signify $1,000. The fact that there were no digits between the two # keystrokes would mean "and no cents." *(Speak the user's language).*

Problem 3. The error message in statement 11 is not precise. It is not clear what "this function" refers to. The problem could be transfers in general, transfers between the two specific accounts, or that the user did not have the $1,000 in the savings account. The system should explicitly state that the user was not allowed to initiate any transfers, thus also avoiding the use of the computer-oriented term "function." The expression "access" is also imprecise as well as being a rather computer-oriented term. An access problem might have been due to system trouble as well as the missing authorization form to allow telephone-initiated transfers. *(Precise and constructive error messages).*

Problem 4. The error message in statement 11 is not constructive. It does not provide any indication of how the user might solve the problem. Users might think that the bank did not want their category of customers to use the transfer facility or that the problem would solve itself if they had more

money in their account. *(Precise and constructive error messages).*

Problem 5. The expression "primary account" in statement 9 is not user-oriented. The system should use user-oriented terms like "checking account." *(Speak the user's language).*

Problem 6. Instead of having the user enter ten digit account numbers, the system could provide the user with a short menu of that user's accounts. There is a much larger risk that the user will make errors when entering a ten digit account number than when entering a single digit menu selection. A menu-based dialogue would probably speed up the dialogue since users would be required to do much less typing and would not need to look up their account numbers. In statement 6, the current design does provide a shortcut by letting the checking account number be the default but this shortcut again involves some risk of errors. Also note that a menu of account names might be difficult to construct if the customer had several accounts of the same type. Assuming that most customers do limit themselves to one of each account, it would still be best to use the menu approach for those customers and stay with the current interface for the difficult customers only: Just because one cannot solve a problem for 100% of the users, one should not skimp out of solving it for, say, the 80% for which a better solution can be found. *(Prevent errors).*

Problem 7. It is very likely that the user will forget to press the # key after having entered menu selections or account numbers. Since the number of digits is predetermined for all user input except for the amount of money, the system in fact does not need a general terminator. The system should only require a # in situations where the input has an indeterminate number of digits and it should then explicitly state the need for this terminator in the prompt. In these few cases, the system could furthermore use a timeout function to give the user a precise and constructive reminder after a certain period of time without any user input, since such a period would normally indicate that the user had finished entering input but had forgotten about the #. *(Prevent errors).*

Problem 8. The feedback in statement 9 with respect to the chosen accounts simply repeats the user's input but ought to restate it instead in simpler and more understandable terms. Instead of listing a ten-digit account number, the feedback message should provide the system's interpretation of the user's input and state something like "from your savings account." By using the name of the account (and by explicitly including the word "your"), the system would increase the user's confidence that the correct account had indeed been specified. *(Provide feedback).*

Minor Problems

Problem 9. The listing of the main menu in statement 1 should reverse the order of the selection number and the function description for each menu item. The current ordering requires users to remember each number as the corresponding description is being spoken since they do not yet know whether they might want to select the function [5]. *(Minimize the user's memory load).*

Problem 10. The most natural order of menu options in this type of system would be a simple numeric order, so the main menu in statement 1 should not skip directly from selection 1 to 3. Users who remember that account transfers were the second option on the list might be inclined to utilize the interrupt facility in the system and simply enter 2 without waiting to hear that the menu choice should have been 3 because there is no option 2 in the system. *(Simple and natural dialogue).*

Problem 11. Feedback on the user's choice of accounts and amounts appears much too late. Normally a lack of feedback would be a "major" problem, but the present design does provide the ** editing facility as well as some feedback (even though it is delayed). *(Provide feedback).*

Problem 12. The options in the accept/cancel menu in statement 9 have been reversed compared to the natural order of the numbers zero and one. Actually it would be possible to achieve some consistency with the rest of the dialogue by using the # key to accept and the * key to cancel. Note that some systems (for instance many British systems) have the reverse convention and use * to indicate the answer yes and # to indicate the answer no. The assignment of meaning to these two keys is more or less arbitrary but should obviously be consistent within the system. The choice between the two meanings of # and * should be made to achieve consistency with the majority of other similar systems in the user's environment. *(Simple and natural dialogue).*

Problem 13. The phrase "account number primary account" in statement 9 is awkward. When referring to an account by name instead of number, the field label "number" should be suppressed. *(Simple and natural dialogue).*

Problem 14. The term "account" in prompts 3 and 5 should be changed to "account number" as the user is required to enter the number. *(Speak the user's language).*

Problem 15. It would probably be better to read out the account numbers one digit at a time instead of using the pairwise grouping in statement 9 since users may well think of their account numbers as grouped differently. The change in feedback method should only apply to the account numbers since it is better to report $1,000 as "one thousand dollars" than as "dollars one zero zero zero." *(Simple and natural dialogue).*

Problem 16. Different words are used for the same concept; "enter" and "press." It is probably better to use the less computer-oriented word "press." *(Consistency).*

The complete voice response system raises several usability issues in addition to the sixteen problems discussed above. One of the most important issues is the voice quality which of course cannot be evaluated in a printed version of the dialogue. Normally one would caution against using the almost identical prompts "Enter account to transfer from/to" (statements 3 and 5) since users could easily confuse them. But the speaker in a voice dialogue can place sufficient emphasis on the words "from" and "to" to make the difference between the prompts obvious.

Results of the Heuristic Evaluation

The BankingSystem in Figure 1 was subjected to heuristic evaluation by three groups of evaluators with varying levels of usability expertise. The first group consisted of 31 computer science students who had completed their first programming course but had no formal knowledge of user interface design principles. These novice evaluators were

	Novice evaluators	*"Regular" specialists*	*"Double" specialists*
Major usability problems:			
1. Error message appears much too late	68%	84%	100%
2. Do not require a dollar amounts to be entered in cents	68%	74%	79%
3. The error message is not precise	55%	63%	64%
4. The error message is not constructive	6%	11%	21%
5. Replace term "primary account" with "checking account"	10%	47%	43%
6. Let users choose accounts from a menu	16%	32%	43%
7. Only require a # where it is necessary	3%	32%	71%
8. Give feedback in form of the name of the chosen account	6%	26%	64%
Average for the major problems	**29%**	**46%**	**61%**
Minor usability problems:			
9. Read menu item description before the action number	3%	11%	71%
10. Avoid the gap in menu numbers between 1 and 3	42%	42%	79%
11. Provide earlier feedback	42%	63%	71%
12. Replace use of 1/0 for accept/reject with #/*	6%	21%	43%
13. Remove the field label "number" when no number is given	10%	32%	36%
14. Change the prompt "account" to "account number"	6%	37%	36%
15. Read numbers one digit at a time	6%	47%	79%
16. Use "press" consistently and avoid "enter"	0%	32%	57%
Average for the minor problems	**15%**	**36%**	**59%**
Average for all the problems	**22%**	**41%**	**60%**

Table 1 *The proportion of evaluators who found each of the sixteen usability problems. "Double" usability specialists had expertise in both usability in general and interfaces to telephone-operated interfaces in particular.*

expected to indicate a worst-case level of performance. Note that they were "novices" with respect to usability but not with respect to computers as such. The second group consisted of 19 "regular" usability specialists, i.e., people with experience in user interface design and evaluation but no special expertise in voice response systems.There is no official certification of usability specialists, but for the purpose of this study, usability specialists were defined as people with graduate degrees and/or several years of job experience in the usability area. The third group consisted of 14 specialists in voice response usability. These "double specialists" had expertise in user interface issues as well as voice response systems and were therefore expected to indicate the best level of heuristic evaluation performance one might hope for.

Table 1 presents the results of the three sets of evaluations and shows that heuristic evaluation was difficult for single evaluators. The above list of usability problems was constructed on the basis of the complete set of evaluations, but no single evaluator found all the problems. Problems 7, 9, 11, 12, 14, and 15 were not included in my own original list of problems but were added after I read the other evaluators' lists. On the other hand, the really catastrophic problems 1, 2, and 3 were found by more than half of the evaluators even in the group without any experience. Just fixing these three problems would improve the interface tremendously.

No group did really well, even though the "double specialists" with both usability expertise and voice response expertise were able to find well over half of the problems on the average. Table 1 indicates that usability specialists are better than people without usability training at finding usability problems and that it helps even more to have usability expertise with respect to the type of user interface being evaluated. The differences between the novices and the regular specialists and between the regular and double specialists are both statistically significant at the $p<.001$ level according to t-tests.

The average performance of individual evaluators may not be acceptable for the use of heuristic evaluation in a usability engineering project, even in the case of the double specialists, but the picture changes when the performance of groups of multiple evaluators is considered. Figure 2 shows the average proportion of the usability problems that would be found by aggregating the sets of problems found by several evaluators. These aggregates were formed in the same way as in previous studies of heuristic evaluation [17]. That is to say, for each group size, a large number of random groups were formed, and for each group, a given usability problem was considered found if at least one member of the group had found it. As can be seen from Figure 2, groups of double and regular usability specialists perform much better than groups of novice evaluators without usability expertise.

For the regular usability specialists, the recommendation from previous work on heuristic evaluation [17] holds in that between three and five evaluators seem necessary to find a reasonably high proportion of the usability problems (here, between 74% and 87%). For the double specialists, however, it is sufficient to use between two and three evaluators to find most problems (here, between 81% and 90%). For the novice evaluators, a group size of fourteen is necessary to find more than 75% of the problems. Using five novice evaluators, which is the upper range of the group size normally recommended for heuristic evaluation, results in the finding of 51% of the usability problems.

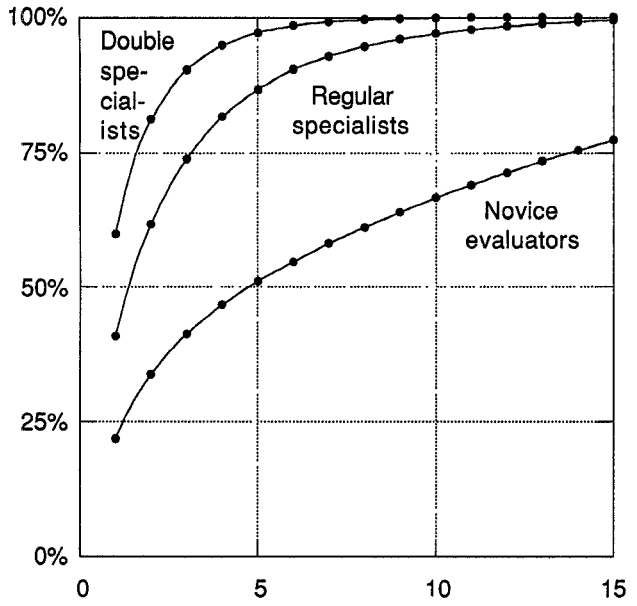


Figure 2 *Average proportion of usability problems found as a function of number of evaluators in a group performing the heuristic evaluation.*

Regular vs. Double Specialists

As mentioned above, the double specialists found significantly more usability problems than did the regular usability specialists. As can be seen from Table 1, the two groups of evaluators actually performed about equally well on many of the usability problems. A large part of the difference in performance is due to the five usability problems for which the probability of being found was thirty percentage points or more higher when the evaluators were voice response usability specialists than when they were regular usability specialists. As outlined below, these five problems were all either specifically related to the use of a telephone as the terminal or were related to the differences between auditory dialogues and screen dialogues.

Problem 9 (read menu item description before the action number) was found by 60% more voice response usability experts than regular usability experts. Even though a similar design issue of whether to list menu selection labels to the left or to the right applies to screen-based menus, the choice would be less crucial for usability. As a matter of fact, screen-based menus are probably better off having the label to the left of the description of the menu item (corresponding to reading the action number before the menu item description) since such a design leads to a uniform, close spacing between the two elements in each line of the menu.

Problem 7 (only require a # where it is necessary) was found by 39% more voice response usability experts than regular usability experts. This problem is much more relevant for telephone-based interfaces than for screen-based interfaces. Actually, the advice to speed up screen-based dialogues by eliminating the need for an enter key wherever possible would probably lead to *less* usable screen interfaces because of the reduced consistency.

Problem 8 (give feedback in form of the name of the chosen account instead of repeating ten digits) was found by 38% more voice response usability experts than regular usability experts. The underlying issue of providing understandable feedback would also apply to screen-based interfaces but the problem would be less serious in such a system because it would be easier for users to understand the ten-digit numbers in their printed form.

Problem 10 (avoid the gap in menu numbers between 1 and 3) was found by 37% more voice response usability experts than regular usability experts. Even though screen-based menus are also more usable when they are sequentially numbered, the numbering is less crucial in the case where the user can see the complete list of numbers simultaneously. A screen-based menu might have a blank line where menu item 2 would normally have been, thus indicating to the user that the number was reserved for a future extension of the system, if that was the reason for omitting the number from the menu. Often, screen menus for non-mouse systems would actually be based on mnemonic characters rather than numbers.

Problem 15 (read numbers one digit at a time) was found by 32% more voice response usability experts than regular usability experts. This problem could only occur in an auditory dialogue and the regular usability specialists would have no prior experience with this exact problem. A similar problem does occur in traditional screen dialogues with respect to the way one should present numbers such as telephone numbers or social security numbers that are normally grouped in a specific way in the user's mind.

These detailed results indicate that the double specialists found more problems, not because they were necessarily better usability specialists in general, but because they had specific experience with usability issues for the kind of user interface that was being evaluated.

In the discussion below of additional factors influencing the finding of usability problems through heuristic evaluation, the results from the "regular" specialists in the BankingSystem evaluation are used since they are the closest to the evaluators used in the other studies that are analyzed.

USABILITY PROBLEM CHARACTERISTICS

Table 2 summarizes six heuristic evaluations. Teledata, Mantel, and the Savings and Transport systems are documented in [17] and the names from that paper are used as headings. For the BankingSystem, the results are given with the "regular" usability specialists as evaluators. The Integrating System was evaluated by "regular" usability specialists and is discussed in [15]. The table only represents those usability problems that were actually found when evaluating the respective interfaces. It is possible that some additional usability problems remain that were not found by anybody, but it is obviously impossible to produce statistics for such problems.

Table 2 also shows three different ways of classifying the usability problems: by severity (i.e., expected impact on the users), by heuristic, and by location in the dialogue. Table 3 then shows the results of an analysis of variance of the finding of the 211 usability problems by single evaluators, with the independent variables being severity, heuristic, and location as well as the system being evaluated and the implementation of its interface. Two implementation categories were used: Teledata, Mantel, and the Banking System were evalu-

Name of interface:	Tele-data		Mantel		Banking System		All paper proto-types		Savings		Trans-port		Inte-grating System		All running systems		All prob-lems	
Number of evaluators:	1	3	1	3	1	3	1	3	1	3	1	3	1	3	1	3	1	3
All problems (211)	.51	.81	.38	.60	.41	.74	.45	.74	.26	.50	.20	.42	.29	.59	.26	.54	**.35**	**.63**
Severity of problem:																		
Major usability problems (59)	.49	.79	.44	.64	.46	.77	.47	.74	.32	.63	.32	.65	.46	.79	.38	.70	**.42**	**.71**
Minor usability problems (152)	.52	.82	.36	.59	.36	.71	.45	.73	.26	.50	.19	.41	.21	.51	.22	.48	**.32**	**.59**
Applicable heuristic:																		
Simple and natural dialogue (51)	.52	.77	.51	.78	.46	.79	.51	.78	.14	.36	.21	.48	.31	.60	.24	.51	**.39**	**.66**
Speak the user's language (35)	.60	.90	.47	.70	.53	.88	.55	.83	.33	.62	.14	.32	.25	.62	.24	.51	**.41**	**.68**
Minimize user memory load (8)	.44	.81	.94	1.00	.11	.30	.48	.73	.26	.62	.15	.39	.27	.57	.24	.54	**.36**	**.63**
Be consistent (33)	.51	.85	.13	.34	.32	.70	.44	.76	.31	.58	.13	.29	.17	.39	.22	.45	**.31**	**.58**
Provide feedback (21)	.60	.87	.68	.96	.45	.79	.58	.87	.39	.72	.48	.85	.39	.69	.40	.72	**.46**	**.77**
Provide clearly marked exits (9)	.19	.50	.03	.09	•	•	.09	.26	.43	.62	.22	.53	•	•	.32	.58	**.20**	**.40**
Provide shortcuts (12)	.39	.78	•	•	•	•	.39	.78	.33	.67	.19	.48	.29	.63	.28	.61	**.29**	**.62**
Good error messages (25)	.42	.73	.33	.62	.37	.63	.38	.66	.23	.46	.33	.53	.27	.66	.25	.48	**.30**	**.56**
Prevent errors (17)	.50	.86	.17	.39	.32	.70	.29	.59	.19	.45	.21	.48	.37	.79	.22	.49	**.25**	**.54**
Where is problem located:																		
A single dialogue element (104)	.58	.85	.42	.66	.40	.75	.49	.77	.26	.53	.22	.45	.30	.59	.26	.53	**.38**	**.66**
Comparison of two elements (43)	.52	.85	.13	.35	.32	.70	.48	.80	.27	.53	.14	.34	.24	.56	.24	.51	**.31**	**.60**
Overall structure of dialogue (18)	.50	.84	•	•	.53	.84	.51	.84	.33	.67	.24	.50	.21	.53	.25	.54	**.35**	**.66**
Something missing (46)	.35	.67	.33	.51	.10	.30	.33	.58	.29	.55	.21	.49	.41	.71	.30	.58	**.31**	**.58**

Table 2 *Proportion of various types of usability problems found in each of the six interfaces discussed in this article, as well as in the collected set of 211 usability problems from all of them. The proportion of problems found is given both when the heuristic evaluation is performed by a single evaluator and when it is performed by aggregating the evaluations from three evaluators. Bullets (•) indicate categories of usability problems that were not present in the interface in question. The total number of usability problems is listed in parentheses for each category.*

ated as paper prototypes, whereas the Savings, Transport, and Integrating Systems were evaluated as running programs.

Even though Table 2 would seem to indicate that paper interfaces are easier to evaluate heuristically than running systems, one cannot necessarily draw that conclusion in general on the basis of the data presented in this paper, since different systems were evaluated in the two conditions. Earlier work on heuristic evaluation [14][17] did speculate that heuristic evaluation might be easier for interfaces with a high degree of persistence that can be pondered at leisure, and it is certainly true that paper prototypes are more persistent than running interfaces.

Table 3 shows that the system being evaluated had a fairly small effect in itself. This would seem to indicate a certain robustness of the heuristic evaluation method, but this result could also be due to the limited range of systems analyzed here. More studies of the application of heuristic evaluation to a wider range of interface styles and application domains will be needed to fully understand which systems are easy to evaluate with heuristic evaluation.

Major vs. Minor Usability Problems

Previous research on heuristic evaluation has pointed out that it identifies many more of the minor usability problems in an interface than other methods do [8]. Indeed, heuristic evaluation picks up minor usability problems that are often not even *seen* in actual user testing. One could wonder to what extent such "problems" should really be accepted as constituting usability problems. I argue that such minor usability problems may very well be real problems even though they are not observable in a user test. For example, inconsistent placement of the same information in different screens or dialog boxes may slow down the user by less than a second [18] and may therefore not be observed in a user test unless an extremely careful analysis is performed on the basis of a large number of videotaped or logged interactions. Such an inconsistency constitutes a usability problem nevertheless, and should be removed if possible. Also note that sub-second slowdowns actually accumulate to causing major costs in the case of highly used systems such as, e.g., those used by telephone company operators.

The top part of Table 2 compares the proportion of the major and the minor usability problems. A usability problem was

	df	Mean Square	*p*	ω^2
Problem severity	1	.842	.001	6.8%
Heuristic used	8	.118	.01	5.0%
Location of problem	3	.047	.37	0.1%
Implementation of interface	1	.747	.07	1.9%
System (nested in Implementation)	4	.123	.03	3.4%
Implementation × Location	3	.159	.02	6.8%
Residual	190	.044		

Table 3 *Analysis of variance for the probability of finding the 211 usability problems when using single evaluators. Other interactions than the one shown are not significant.*

ω^2 *indicates relative effect sizes in terms of proportion of total variance accounted for.*

defined as "major" if it had serious potential for confusing users or causing them to use the system erroneously. Note that the term "serious" was used to denote this category of usability problems in earlier work [12]. Given that the usability problems were found by heuristic evaluation and not by user testing, this classification can only reflect a considered judgment, since no measurement data exists to prove the true impact of each problem on the users. For the Teledata, Mantel, Savings, and Transport interfaces, the major/minor classification was arrived at by two judges with a small number of disagreements resolved by consensus, and for the Banking System a single judge was used. For the Integrating System, the mean severity classification from eleven judges was used. The simple classification of usability problems into only two severity levels was chosen because of this need to rely on a judgment; it was mostly fairly easy to decide which severity category to use for any given usability problem. See [9] and [15] for further discussions of severity ratings.

It is apparent from Table 2 that heuristic evaluation tends to find a higher proportion of the major usability problems than of the minor, and Table 3 indicates that the difference is statistically significant ($p<.001$) and one of the two largest effects identified in the table. Intuitively, one might even have gone as far as to expect the evaluators performing the heuristic evaluations to focus only on the major usability problems to the exclusion of the minor ones, but the results indicate that this is not the case since they find many more minor than major problems in absolute numbers (8.1 vs. 4.1 per system on the average). So the evaluators pay relatively more attention to the major problems without neglecting the minor ones.

Since the interfaces have many more minor than major problems, the minor problems will obviously dominate any given heuristic evaluation, even though the probability of being found is greater for the major problems. Usability engineers therefore face the task of prioritizing the usability problems to make sure that more time is spent on fixing the major problems than on fixing the minor problems.

Effect of the Individual Heuristics

Since heuristic evaluation is based on judging interfaces according to established usability principles, one might expect that problems violating certain heuristics would be easier to find than others. Table 3 indicates a significant and fairly large effect for heuristic. Even so, Table 2 shows that there are few systematic trends with respect to some heuristics being easier.

Considering all the 211 usability problems as a whole, Table 2 shows that usability problems have about the same probability of being found in a heuristic evaluation with the recommended three evaluators for most of the heuristics. Seven of the nine heuristics score in the interval from 54–68%, with the "good error messages" and "prevent errors" heuristics being slightly more difficult than the others. The only truly difficult heuristic is "provide clearly marked exits" (scoring 40%). The practical consequence from this result is that one might "look harder" for usability problems violating the "provide clearly marked exits" heuristic. For example, one could run a user test with a specific focus on cases where the users got stuck. One could also study user errors more closely in order to compensate for the relative difficulty of applying the two error-related heuristics, especially since problems related to user errors are likely to prove especially costly if the system were to be released with these problems still in place.

A contrast analysis of significance based on an analysis of variance for three evaluators confirms that usability problems classified under the "good error messages," "prevent errors," and "provide clearly marked exits" heuristics are more difficult to find than usability problems classified under one of the other six heuristics, with $p=.0006$.

Location of Problems in Dialogue

Even though the specific usability heuristic used to classify the usability problems had some impact on the evaluators' ability to find the problems, it might also be the case that other systematic differences between the problems can help explain why some problems are easier to find than others. Since heuristic evaluation is a process in which the evaluators search for usability problems, it seems reasonable to consider whether the circumstances under which the problems could be located have any influence.

The bottom part of Table 2 shows the result of considering four different possible locations of usability problems. The first category of problems are those that are located in a *single* dialogue element. An example of this category of usability problem is Problem 2 (do not require a dollar amount to be entered as cents) in the telephone operated interface analyzed earlier in this article. To find single-location problems by heuristic evaluation, the evaluator only needs to consider each interface element in isolation and judge that particular dialog box, error message, menu, etc.

The second category consists of usability problems that require the evaluator to *compare* two interface elements. This will typically be consistency problems where each interface element is fine when seen in isolation but may lead to problems when used together. An example from the BankingSystem is Problem 16 (both "press" and "enter" are used to denote the same concept).

The third category contains the usability problems that are related to the overall *structure* of the dialogue. An example from the BankingSystem is Problem 7 (only require a # where it is necessary). Another example would be the need to unify the navigation system for a large menu structure. These problems require the evaluator to get a grasp of the overall use of the system.

The final category of usability problems are those that cannot be seen in any current interface element but denote *missing* interface elements that ought to be there. An example from the BankingSystem is Problem 4 (the error message should have a constructive message appended). Note that the issue here is not that the current error message is poorly worded (that is easy to find and belongs in the category of single-location problems) but that the message ought to be supplemented with an additional element.

As can be seen from Table 3, the difference between the four location categories is not statistically significant. However, the interaction effect between location category and interface implementation *is* significant and has one of the two largest effect sizes in the table. As shown in Table 2, problems in the category "something missing" are slightly easier to find than other problems in running systems but much harder to find

than other problems in paper prototypes. This finding corresponds to an earlier, qualitative, analysis of the usability problems that were harder to find in a paper implementation than in a running system [14]. Because of this difference, one should look harder for missing dialogue elements when evaluating paper mockups.

A likely explanation of this phenomenon is that evaluators using a running system may tend to get stuck when needing a missing interface element (and thus notice it), whereas evaluators of a paper "implementation" just turn to the next page and focus on the interface elements found there.

CONCLUSIONS

Usability specialists were much better than those without usability expertise at finding usability problems by heuristic evaluation. Furthermore, usability specialists with expertise in the specific kind of interface being evaluated did much better than regular usability specialists without such expertise, especially with regard to certain usability problems that were unique to that kind of interface.

Previous results [17] with respect to the improvement in heuristic evaluation performance as groups of evaluators are aggregated were replicated in the new study reported above, and the general recommendation of using groups of 3–5 evaluators also held for the regular usability specialists in this study. For double specialists, a smaller group size can be recommended, since only two to three such evaluators were needed to find most problems. Of course, the actual number of evaluators to use in any particular project will depend on a trade-off analysis on the basis of curves like Figure 2 and the cost (financial or otherwise) of leaving usability problems unfound.

Major usability problems have a higher probability than minor problems of being found in a heuristic evaluation, but about twice as many minor problems are found in absolute numbers. Problems with the lack of clearly marked exits are harder to find than problems violating the other heuristics, and additional efforts should therefore be taken to identify such usability problems. Also, usability problems that relate to a missing interface element are harder to find when an interface is evaluated in a paper prototype form.

The results in this article provide means for improving the contribution of heuristic evaluation to an overall usability engineering effort. The expertise of the staff performing the evaluation has been seen to matter, and specific shortcomings of the methods have been identified such that other methods or additional efforts can be employed to alleviate them and find more of the usability problems that are hard to find by heuristic evaluation.

ACKNOWLEDGMENTS

The author would like to thank Jan C. Clausen, Heather Desurvire, Dennis Egan, Anker Helms Jørgensen, Clare-Marie Karat, Tom Landauer, Rolf Molich, and Robert W. Root for helpful comments on previous versions of the manuscript. The four studies reported in [17] were conducted by the author and Rolf Molich who also participated in the classification of usability problems as major or minor and in relating the problems to the heuristics. The further analyses and conclusions on the basis of this and other data as reported here reflect the views of the author of the present paper only.

REFERENCES

1. Bias, R. Walkthroughs: Efficient collaborative testing. *IEEE Software* **8**, 5 (September 1991), 94–95.
2. Carroll, J.M. Infinite detail and emulation in an ontologically minimized HCI. *Proc. ACM CHI'90* (Seattle, WA, 1–5 April 1990), 321–327.
3. Carroll, J.M., Kellogg, W.A., and Rosson, M.B. The task-artifact cycle. In Carroll, J.M. (Ed.), *Designing Interaction: Psychology at the Human–Computer Interface*. Cambridge University Press, Cambridge, U.K., 1991. 74–102.
4. Desurvire, H., Lawrence, D., and Atwood, M. Empiricism versus judgement: Comparing user interface evaluation methods on a new telephone-based interface. *ACM SIGCHI Bulletin* **23**, 4 (October 1991), 58–59.
5. Engelbeck, G., and Roberts, T.L. The effect of several voice-menu characteristics on menu selection performance. *Behaviour & Information Technology* **in press**.
6. Gould, J.D., and Boies, S.J. Speech filing—An office system for principals. *IBM Systems Journal* **23**, 1 (1984), 65–81.
7. Halstead-Nussloch, R. The design of phone-based interfaces for consumers. *Proc. ACM CHI'89* (Austin, TX, 30 April–4 May 1989), 347–352.
8. Jeffries, R., Miller, J.R., Wharton, C., and Uyeda, K.M. User interface evaluation in the real world: A comparison of four techniques. *Proc. ACM CHI'91* (New Orleans, LA, 27 April–2 May 1991), 119–124.
9. Karat, C.-M., Campbell, R., Fiegel, T. Comparisons of empirical testing and walkthrough methods in user interface evaluation. *Proc. ACM CHI'92* (Monterey, CA, 3–7 May 1992).
10. Kellogg, W.A. Qualitative artifact analysis. *Proc. INTERACT'90 3rd IFIP Conf. Human–Computer Interaction* (Cambridge, U.K., 27–31 August 1990), 193–198.
11. Lewis, C., Polson, P., Wharton, C., and Rieman, J. Testing a walkthrough methodology for theory-based design of walk-up-and-use interfaces. *Proc. ACM CHI'90* (Seattle, WA, 1–5 April 1990), 235–241.
12. Molich, R., and Nielsen, J. Improving a human-computer dialogue. *Communications of the ACM* **33**, 3 (March 1990), 338–348.
13. Nielsen, J. Usability engineering at a discount. In Salvendy, G., and Smith, M.J. (Eds.), *Designing and Using Human–Computer Interfaces and Knowledge Based Systems*, Elsevier Science Publishers, Amsterdam, 1989. 394–401.
14. Nielsen, J. Paper versus computer implementations as mockup scenarios for heuristic evaluation. *Proc. INTERACT'90 3rd IFIP Conf. Human–Computer Interaction* (Cambridge, U.K., 27–31 August 1990), 315–320.
15. Nielsen, J. Applying heuristic evaluation to a highly domain-specific interface. *Manuscript submitted for publication.*
16. Nielsen, J. *Usability Engineering*. Academic Press, San Diego, CA, 1992.
17. Nielsen, J., and Molich, R. Heuristic evaluation of user interfaces. *Proc. ACM CHI'90* (Seattle, WA, 1–5 April 1990), 249–256.
18. Teitelbaum, R.C., and Granda, R.E. The effects of positional constancy on searching menus for information. *Proc. ACM CHI'83* (Boston, MA, 12–15 December 1983), 150–153.
19. Wharton, C., Bradford, J., Jeffries, R., and Franzke, M. Applying cognitive walkthroughs to more complex interfaces: Experiences, issues, and recommendations. *Proc. ACM CHI'92* (Monterey, CA, 3–7 May 1992).

HEURISTIC EVALUATION OF USER INTERFACES

Jakob Nielsen and *Rolf Molich*

Technical University of Denmark
Department of Computer Science
DK-2800 Lyngby Copenhagen
Denmark
datJN@NEUVM1.bitnet

Baltica A/S
Mail Code B22
Klausdalsbrovej 601
DK-2750 Ballerup
Denmark

ABSTRACT

Heuristic evaluation is an informal method of usability analysis where a number of evaluators are presented with an interface design and asked to comment on it. Four experiments showed that individual evaluators were mostly quite bad at doing such heuristic evaluations and that they only found between 20 and 51% of the usability problems in the interfaces they evaluated. On the other hand, we could aggregate the evaluations from several evaluators to a single evaluation and such aggregates do rather well, even when they consist of only three to five people.
KEYWORDS: *Usability evaluation, early evaluation, usability engineering, practical methods.*

INTRODUCTION

There are basically four ways to evaluate a user interface: *Formally* by some analysis technique, *automatically* by a computerized procedure, *empirically* by experiments with test users, and *heuristically* by simply looking at the interface and passing judgement according to ones own opinion. Formal analysis models are currently the object of extensive research but they have not reached the stage where they can be generally applied in real software development projects. Automatic evaluation is completely infeasible except for a few very primitive checks. Therefore current practice is to do empirical evaluations if one wants a good and thorough evaluation of a user interface. Unfortunately, in most practical situations, people actually *do not* conduct empirical evaluations because they lack the time, expertise, inclination, or simply the tradition to do so. For example, Milsted et al. [1989] found that only 6% of Danish companies doing software development projects used the thinking aloud method and that nobody used *any* other other empirical or formal evaluation methods.

In real life, most user interface evaluations are heuristic evaluations but almost nothing is known about this kind of evaluation since it has been seen as inferior by most researchers. We believe, however, that a good strategy for improving usability in most industrial situations is to study those usability methods which are likely to see practical use [Nielsen 1989]. Therefore we have conducted the series of experiments on heuristic evaluation reported in this paper.

HEURISTIC EVALUATION

As mentioned in the introduction, heuristic evaluation is done by looking at an interface and trying to come up with an opinion about what is good and bad about the interface. Ideally people would conduct such evaluations according to certain rules, such as those listed in typical guidelines documents. Current collections of usability guidelines [Smith and Mosier 1986] have on the order of one thousand rules to follow, however, and are therefore seen as intimidating by developers. Most people probably perform heuristic evaluation on the basis of their own intuition and common sense instead.

We have tried cutting the complexity of the rule base by two orders of magnitudes by relying on a small set of heuristics such as the nine basic usability principles from [Molich and Nielsen 1990] listed in Table 1. Such smaller sets of principles seem more suited as the basis for practical heuristic evaluation. Actually the use of very

Simple and natural dialogue
Speak the user's language
Minimize user memory load
Be consistent
Provide feedback
Provide clearly marked exits
Provide shortcuts
Good error messages
Prevent errors

Table 1. *Nine usability heuristics (discussed further in [Molich and Nielsen 1990]).*

complete and detailed guidelines as checklists for evaluations might be considered a formalism, especially when they take the form of interface standards.

We have developed this specific list of heuristics during several years of experience with teaching and consulting about usability engineering [Nielsen and Molich 1989]. The nine heuristics can be presented in a single lecture and explain a very large proportion of the problems one observes in user interface designs. These nine principles correspond more or less to principles which are generally recognized in the user interface community, and most people might think that they were "obvious" if it was not because the results in the following sections of this paper show that they are difficult to apply in practice. The reader is referred to [Molich and Nielsen 1990] for a more detailed explanation of each of the nine heuristics.

EMPIRICAL TEST OF HEURISTIC EVALUATION

To test the practical applicability of heuristic evaluation, we conducted four experiments where people who were not usability experts analyzed a user interface heuristically. The basic method was the same in all four experiments: The evaluators ("subjects") were given a user interface design and asked to write a report pointing out the usability problems in the interface as precisely as possible. Each report was then scored for the usability problems that were mentioned in it. The scoring was done by matching with a list of usability problems developed by the authors. Actually, our lists of usability problems had to be modified after we had made an initial pass through the reports, since our evaluators in each experiment discovered some problems which we had not originally identified ourselves. This shows that even usability experts are not perfect in doing heuristic evaluations.

Scoring was liberal to the extent that credit was given for the mentioning of a usability problem even if it was not described completely.

Table 2 gives a short summary of the four experiments which are described further in the following.

Experiment (short name)	*No. Evaluators*	*Total Known Usability Problems*	*Average Problems Found*
Teledata	37	52	51%
Mantel	77	30	38%
Savings	34	48	26%
Transport	34	34	20%

Table 2. Summary of the four experiments.

Experiment 1: Teledata

Experiment 1 tested the user interface to the Danish videotex system, Teledata. The evaluators were given a set of ten screen dumps from the general search system and from the Scandinavian Airlines (SAS) subsystem. This means that the evaluators did not have access to a "live" system, but in many situations it is realistic to want to conduct a usability evaluation in the specification stage of a software development process where no running system is yet available.

The evaluators were 37 computer science students who were taking a class in user interface design and had had a lecture on our evaluation heuristics before the experiment. The interface contained a total of 52 known usability problems.

Experiment 2: Mantel

For experiment 2 we used a design which was constructed for the purpose of the test. Again the evaluators had access only to a written specification and not to a running system. The system was a design for a small information system which a telephone company would make available to its customers to dial in via their modems to find the name and address of the subscriber having a given telephone number. This system was called "Mantel" as an abbreviation of our hypothetical telephone company, Manhattan Telephone (neither the company nor the system has any relation to any existing company or system). The entire system design consisted of a single screen and a few system messages so that the specification could be contained on a single page.

The design document used for this experiment is reprinted as an appendix to [Molich and Nielsen 1990] which also gives a complete list and in-depth explanation of the 30 known usability problems in the Mantel design.

The evaluators were readers of the Danish *Computerworld* magazine where our design was printed as an exercise in a contest. 77 solutions were mailed in, mostly written by industrial computer professionals. Our main reason for conducting this experiment was to ensure that we had data from real computer professionals and not just from students. We should note that these evaluators did not have the (potential) benefit of having attended our lecture on the usability heuristics.

Experiments 3 and 4: Two Voice Response Systems: "Savings" and "Transport"

Experiments 3 and 4 were conducted to get data from heuristic evaluations of "live" systems (as opposed to the specification-only designs in experiments 1 and 2). Both experiments were done with the same group of 34 computer science students as evaluators. Again, the students were taking a course in user interface design and were given a lecture on our usability heuristics, but there was no overlap between the group of evaluators in these experiments and the group from experiment 1.

Both interfaces were "voice response" systems where users would dial up an information system from a touch tone telephone and interact with the system by pushing buttons on the 12-key keypad. The first system was run by a large Savings Union to give their customers information about their account balance, current foreign currency exchange rates, etc. This interface is referred to as the "Savings" de-

sign in this article and it contained a total of 48 known usability problems. The second system was used by the municipal public transportation company in Copenhagen to provide commuters with information about bus routes. This interface is referred to as the "Transport" design and had a total of 34 known usability problems.

There were four usability problems which were related to inconsistency across the two voice response systems. Since the two systems are aimed at the same user population in the form of the average citizen and since they are accessed through the same terminal equipment, it would improve their collective usability if they both used the same conventions. Unfortunately there are differences, such as the use of the square[1] key. In the Savings system, it is an end-of-command control character, while it is a command key for the "return to the main menu" command in the Transport system which does not use an end-of-command key at all. The four shared inconsistency problems have been included in the count of usability problems for both systems.

Since the same evaluators were used for both voice response experiments, we can compare the performance of the individual evaluators. In this comparison, we have excluded the four consistency problems discussed above which are shared among the two systems. A regression analysis of the two sets of evaluations is shown in Figure 1 and indicates a very weak correlation between the performance of the evaluators in the two experiments (R^2=0.33, p<0.01). So while some people *are* better than others at doing heuristic evaluation of user interfaces, this tendency is not very strong. We do not have enough evidence to form a firm conclusion but it seems that it might be the case that there is very little consistency in the ability of evaluators to find usability problems. The two evaluations compared in Figure 1 concerned quite similar interfaces (both were voice response systems), and it would be a plausible hypothesis that evaluators would perform even less consistently in evaluations of more varied systems.

We should note that the evaluators in these two experiments all had the same level of usability expertise. Even though we do not have formal evidence to show this, we do believe that usability experts will be better at heuristic evaluation than average computer professionals. It is likely that experience in usability and empirical user tests provides a good background for recognizing and conceptualizing usability problems. With regard to the latter, expertise in *running* user tests would probably not be as much help as the observations of actual user behavior made by the experienced tester over the years.

[1] This key is also sometimes called the "pound key". In fact one of the inconsistency problems was that this single key had two different names in the two systems (*firkant* and *rude*, respectively, in Danish).

THE USABILITY PROBLEMS

We have already mentioned a usability problem related to the "consistency" rule in the description of experiments 3 and 4. A few other examples of usability problems are:

- The Mantel system overwrites the telephone number entered by the user so that it is no longer visible when the name and address of the corresponding subscriber is displayed (found by 95%).
- The Transport system shifts from reading submenus to reading the main menu without any pause or indication that the user is moved to another level of menu (found by 62%).
- The error message "UNKNOWN IP" in Teledata (where IP stands for information provider) can be made much more readable (found by 54%).
- Users who do not have the printed user's guide will never learn that the Savings system has an online help facility (found by 35%).
- The key to accessing certain information in the Transport system is the transport company's internal departmental organization instead of the bus numbers known by the public (found by 12%).

The validity of these usability problems is an important question: Will they in fact present problems to real users, and to what degree do they constitute the complete set of usability problems? We have not conducted traditional empirical usability tests to measure this. On the other hand, we do have two arguments in support for the validity of the problems as usability problems. The first argument is simply that most of these design issues are "obviously" problems according to established knowledge in the usability field. The second, and perhaps more convincing argument is that the very method of our experiments actually forms a kind of empirical support for the usability problems. For

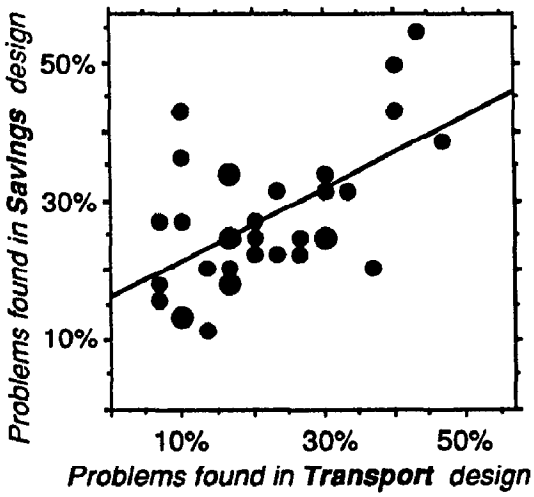


Figure 1. *Scatterplot of the proportion of usability problems found by the same evaluators in two different interfaces. The regression line has R^2=0.33 and shows that there is only a very weak correlation between the evaluators' performance in the two experiments.*

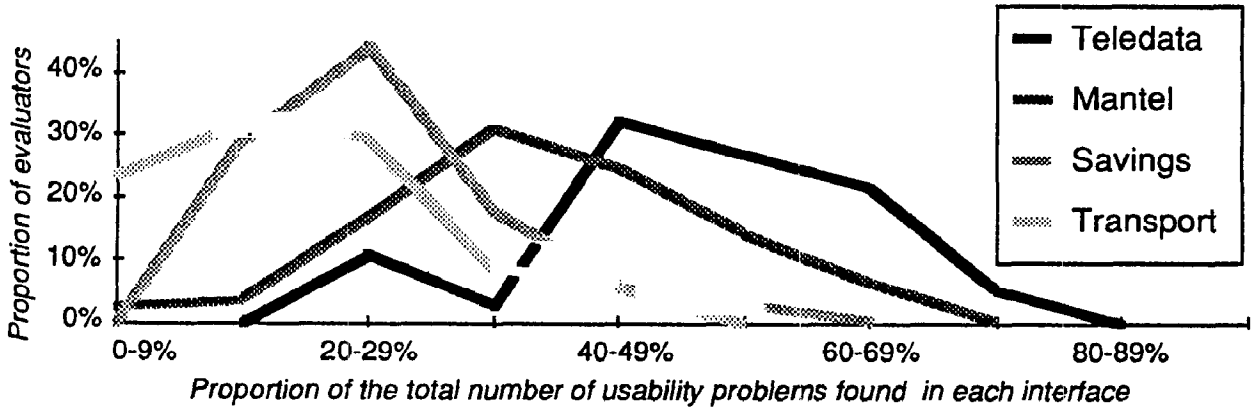


Figure 2. *Distribution for each of the four experiments of the number of usability problems found by the evaluators (expressed as percent of the total number of problems in each interface to enable comparisons).*

each system we have had at least 34 people work their way through the interface. If we view these people as experimental subjects rather than as evaluators, we realize that it is very unlikely that any of the systems would have had any major usability problem which did not bother some of these subjects enough to complain about it in their report.

In spite of these arguments, it is always impossible to know for sure whether one has found every single usability problem in an interface. It might be that the next subject would stumble over something new. Therefore we have stated for each experiment the "known" number of usability problems, and the statistics in the following sections are based on this number of known problems.

Furthermore, the usability problems of an interface do not form a fixed set in real life. For any actual use of a system by real users in a real context, only some of its potential weaknesses will surface as problems. Some aspects of a design might never bother a particular user and could therefore not be said to be "problems" as far as that user is concerned. Even so, we will still consider a design item as a usability problem if it could be expected to bother some users during some reasonable use of the system. The decision whether or not to remove the problem in a redesign should then be based on a judgement of the number of users it impacts and a trade-off analysis of whether removing it would reduce the efficiency of use or other desirable usability parameters for other users. One can only get the option to make this judgement and trade-off analysis, however, if one has identified the usability problem in the first place.

A weakness of our approach is that we only looked at individual usability "problems" in the phase of a development process where one has completed the overall design and needs to polish it. It would also be interesting to consider more holistic evaluations of entire interfaces such as those that would be required to select which of two competing products to purchase or which of two completely different design approaches to pursue. It is likely, however, that a different set of techniques will be needed for that kind of evaluation.

EVALUATION RESULTS

The most basic result from the four experiments is that heuristic evaluation is difficult. The average proportion of usability problems found was 51%, 38%, 26%, and 20% in the four experiments respectively. So even in the best case only half of the problems were found, and the general case was rather poor. Actually, even these numbers are not all that bad. Even finding *some* problems is of course much better than finding *no* problems, and one could supplement the heuristic method with other usability engineering methods to increase the total number of problems found.

Figure 2 shows the distribution of the number of problems found in each of the four experiments. We can see that the distributions as expected mostly have a shape like the normal distribution, even though the curve for the Transport experiment is somewhat skewed. In other words, most evaluators do about average, a few do very well, and a few do rather badly.

Table 3 presents information related to individual differences in the performance of evaluators. First, the number of usability problems found is expressed in percent of the

	N	Min %	Max %	D_1 %	D_9 %	Q_1 %	Q_3 %	Max/Min	D_9/D_1	Q_3/Q_1
Teledata	37	22.6	74.5	26.6	67.9	43.2	58.5	3.3	2.6	1.4
Mantel	77	0 [6.7]	63.3	23.3	53.3	30.0	46.7	∞ [9.4]	2.3	1.6
Savings	34	10.4	52.1	14.4	39.8	18.8	31.3	5.0	2.8	1.7
Transport	34	6.7	46.7	8.8	35.6	11.8	26.5	7.0	4.0	2.2
Average		**13.2**	**59.3**	**18.3**	**49.2**	**26.0**	**40.8**	**5.1**	**2.9**	**1.7**

Table 3. *Individual differences in evaluators' ability to find usability problems.*

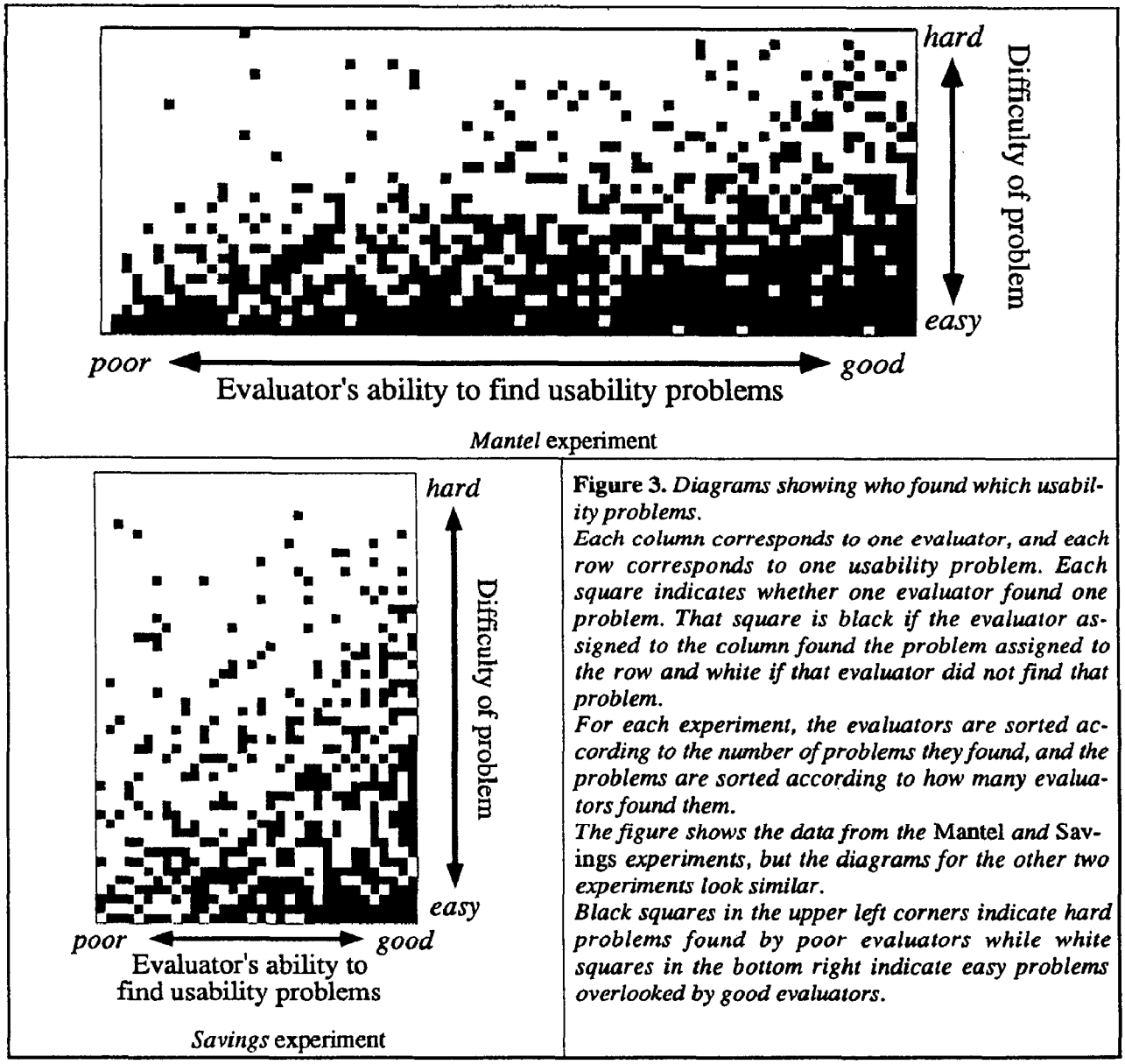


Figure 3. *Diagrams showing who found which usability problems.*
Each column corresponds to one evaluator, and each row corresponds to one usability problem. Each square indicates whether one evaluator found one problem. That square is black if the evaluator assigned to the column found the problem assigned to the row and white if that evaluator did not find that problem.
For each experiment, the evaluators are sorted according to the number of problems they found, and the problems are sorted according to how many evaluators found them.
The figure shows the data from the Mantel *and* Savings *experiments, but the diagrams for the other two experiments look similar.*
Black squares in the upper left corners indicate hard problems found by poor evaluators while white squares in the bottom right indicate easy problems overlooked by good evaluators.

total number of usability problems in each interface. For each of the five experiments, the table then lists the proportion of problems found by the worst and best evaluator, the first and ninth decile, and first and third quartile, as well as the ratios between these values. In the Mantel experiment, one of the evaluators did not find any problems at all, so the table also lists the problems found by the second worst evaluator. The Mantel experiment has been excluded from the calculation of the averages of the minimums and of the max/min ratios.

We see that the individual differences correspond to the Q_3/Q_1 ratios of about 2 listed by Egan [1988] for text editing but are lower than the ratios of 2 to 4 listed for information search and for programming. They correspond closely to the Q_3/Q_1 ratio of 1.8 for time needed to learn HyperCard programming [Nielsen 1990] by the same category of computer science students as those used in three of the four experiments.

We see from Tables 2 and 3 that some systems are easier to evaluate heuristically than others. One interesting trend from Table 3 is that the individual differences between evaluators are larger the more difficult the interface is to evaluate. Table 2 further shows that the voice response systems were especially hard to evaluate. The problem with heuristic evaluation of voice interfaces is that they have an extremely low persistence [Nielsen 1987] because all system messages are gone as soon as they are uttered. This again means that evaluators get no chance to ponder details of the interface design at their leisure.

In general, there were rather few false positives in the form of evaluators stating that something was a usability problem when we would not classify it as such. Therefore we have not conducted a formal analysis of false positives. For a practical application of heuristic evaluation, false positives might present a problem to the extent that one evaluator's finding of a false positive could sidetrack the discus-

sion in a development group. Our experience is that a given false positive normally is not found by more than a single evaluator, so the other members of the development group should be able to convince the finder of the false positive that it is not a real usability problem. If not, then an empirical test could serve as the ultimate arbiter. We would in general recommend that one does not rely exclusively on heuristic evaluation during the usability engineering process. Such methods as thinking aloud should be used to supplement the heuristic evaluation results in any case.

We should note that we have only tested heuristic evaluation of fairly small-scale interfaces. We do not know what happens during the heuristic evaluation of much larger interface designs. Furthermore, we studied evaluations of complete designs in the form of paper prototypes or running systems. It is also of interest what happens during the "inner loop" of design [Newell and Card 1985] where a designer rapidly evaluates various alternative subdesigns before they are finalized in a complete design. It is likely that such evaluations are often heuristic in nature, so some of the same results may apply.

AGGREGATED EVALUATIONS

Figure 2 and Table 3 show that some evaluators do better than others. One might have supposed that the difference in performance between evaluators was due to an inherent rank ordering of the difficulty of finding the usability problems, such that a "good" evaluator would be able to find all the easy problems found by a "poor" evaluator as well as some additional, harder problems. Figure 3 shows, however, that this is not the case. Even poor evaluators can sometimes find hard problems as indicated by the black squares in the upper left part of the diagrams. And good evaluators may sometimes overlook easy problems as indicated by the white squares in the lower right part of the diagrams. In other words, the finding of usability problems does not form a perfect cumulative scale (a Guttman[2] scale [Guttman 1944]).

Because of this phenomenon, we have the potential for dramatically improving the overall result by forming *aggregates* of evaluators since the "collected wisdom" of several evaluators is not just equal to that of the best evaluator in the group. Aggregates of evaluators are formed by having several evaluators conduct a heuristic evaluation and then collecting the usability problems found by each of them to form a larger set.

For this aggregation process to work, we have to assume that there is some authority that is able to read through the reports from the individual evaluators and recognize the usability problems from each report. This authority could be a usability expert or or it could be the group itself during a meeting of the evaluators. We have not tested this assumption empirically but it seems reasonable for the kind of usability problems discussed in this paper since they are of a nature where they are "obvious" as soon as somebody has pointed them out.

Our experience from conducting the four experiments and discussing them with the evaluators indicates that people are usually willing to concede that something is a usability problem when it is pointed out to them by others. At least for the kind of usability problems considered in this paper, the main difficulty lies in finding them in the first place, not in agreeing on the aggregated list.

On the basis of our data showing which evaluators found which usability problems, we have constructed hypothetical aggregates of varying sizes to test how many problems such aggregates would theoretically find. The aggregates were not formed in a real project but given our assumption of a perfect authority to form the conclusions, that should make no difference. For each of our four experiments, aggregates were formed by choosing the number of people in the aggregate randomly from the total set of evaluators in that experiment. For each experiment, it would of course have been possible to select an optimal aggregate of the better evaluators but in a real company one would not have that luxury. Normally one would have to use whatever staff was available, and that staff would have been hired on the basis of many other qualifications than their score in heuristic evaluation experiments. And in any case, Figure 1 indicates that people who are good evaluators in one experiment may not be all that good in the next experiment.

Figure 4 shows the results from selecting random aggregates of evaluators. The figure shows the average number of usability problems found by each size of aggregate. These averages were calculated by a Monte Carlo technique where we selected between five and nine thousand

[2] The evaluations do approximate a Guttman scale with an average Guttman reproducibility coefficient R = 0.85 (coefficients ranging from 0.82 to 0.87). The average minimal marginal reproducibility, MMR is 0.80 (ranging from 0.79 to 0.82), however, indicating that the scale is not truly unidimensional and cumulative since the coefficient of scalability is only 0.06. The Guttman coefficient indicates the degree to which the data follows a unidimensional cumulative scale, with a value of 1 indicating a perfect scale. The Guttman coefficient of 0.85 shows that only 15% of the data deviates from that expected of such a perfect scale. But the minimal marginal reproducibility indicates the degree to which the individual values could be predicted from the average values even disregarding potential scaling properties. From knowing e.g. that a certain usability problem was only found by 20% of the evaluators, we would be able to correctly predict 80% of the data for that problem without taking that evaluators general problem-finding abilities into account by just predicting for each evaluator that he or she would not find the problem. So the assumption of strict ordering only gains us an improvement from 80% to 85%, indicating that it has poor explanatory powers. In any case, it is the deviation of 15% from the Guttman scale which allows us to form the aggregates we discuss here.

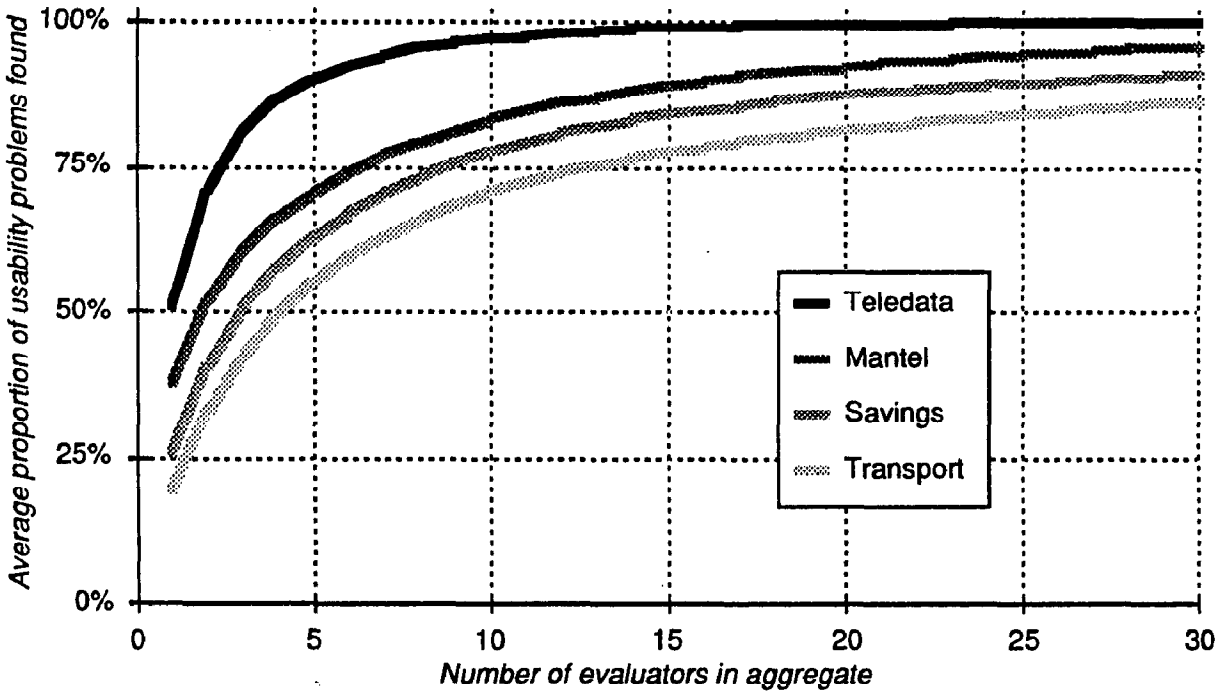


Figure 4. *Proportion of usability problems found by aggregates of size 1 to 30.*

random aggregates for each aggregate size and experiment. Table 4 gives the exact numbers for selected sizes of aggregates.

It is apparent from Figure 4 that the curves for the four experiments have remarkably similar shapes. Each curve rises drastically in the interval from one evaluator to about five evaluators, it then flattens out somewhat around the interval from five to ten evaluators, and the point of diminishing returns seems to have been reached at aggregates of about ten evaluators. It is interesting to see that even for the Transport interface which was the hardest to analyze, aggregates of five evaluators are still able to find more than half of the usability problems. In general, we would expect aggregates of five evaluators to find about two thirds of the usability problems which is really quite good for an informal and inexpensive technique like heuristic evaluation.

For the aggregated evaluation to produce better results than the individual evaluations, it is likely that the evaluators should do their initial evaluations independently of each other and only compare results *after* each of them has looked at the design and written his/her evaluation report. The reason we believe this is that evaluators working together in the initial evaluation phase might tend to bias each other towards a certain way of approaching the analysis and therefore only discover certain usability problems. It is likely that the variety in discovered usability problems apparent in our experiments would have been smaller if the evaluators had worked in groups. And it is of course the variety which is the reason for the improvement one gets from using aggregates of evaluators.

Aggregate:	*1*	*2*	*3*	*5*	*10*
Teledata	51%	71%	81%	90%	97%
Mantel	38%	52%	60%	70%	83%
Savings	26%	41%	50%	63%	78%
Transport	20%	33%	42%	55%	71%

Table 4. *Average proportion of usability problems found in each of the four interfaces for various sized aggregates of evaluators.*

CONCLUSIONS

This study shows that heuristic evaluation is difficult and that one should not rely on the results of having a single person look at an interface. The results of a heuristic evaluation will be much better if you have several people conduct the evaluation, and they should probably do so independently of each other. The number of usability results found by aggregates of evaluators grows rapidly in the interval from one to five evaluators but reaches the point of diminishing returns around the point of ten evaluators. We recommend that heuristic evaluation is done with between three and five evaluators and that any additional resources are spent on alternative methods of evaluation.

Major advantages of heuristic evaluation are:
- It is cheap.
- It is intuitive and it is easy to motivate people to do it.
- It does not require advance planning.
- It can be used early in the development process.

A disadvantage of the method is that it sometimes identifies usability problems without providing direct suggestions for how to solve them. The method is biased by the current mindset of the evaluators and normally does not generate breakthroughs in the evaluated design.

ACKNOWLEDGEMENTS

The authors would like to thank Jan C. Clausen, John Schnizlein, and the anonymous *CHI'90* referees for helpful comments.

REFERENCES

1. Egan, D.E. Individual differences in human-computer interaction. In: M. Helander (Ed.): *Handbook of Human-Computer Interaction*. Elsevier Science Publishers, Amsterdam, 1988, pp. 543-568.
2. Guttman, L. A basis for scaling qualitative data. *American Sociological Review* **9** (1944), 139-150.
3. Milsted, U., Varnild, A., and Jørgensen, A.H. Hvordan sikres kvaliteten af brugergrænsefladen i systemudviklingen ("Assuring the quality of user interfaces in system development," in Danish). *Proc. NordDATA'89 Joint Scandinavian Computer Conference* (Copenhagen, Denmark, 19-22 June 1989), 479-484.
4. Molich, R. and Nielsen, J. Improving a human-computer dialogue: What designers know about traditional interface design. *Communications of the ACM* **33**, 3 (March 1990).
5. Newell, A. and Card, S.K. The prospects for psychological science in human-computer interaction. *Human-Computer Interaction* **1**, 3 (1985), 209-242.
6. Nielsen, J. Classification of dialog techniques: A CHI+GI'87 workshop, Toronto, April 6, 1987. *ACM SIGCHI Bulletin* **19**, 2 (October 1987), 30-35.
7. Nielsen, J. Usability engineering at a discount, in Salvendy, G. and Smith, M.J. (Eds.): *Designing and Using Human-Computer Interfaces and Knowledge Based Systems*. Elsevier Science Publishers, Amsterdam 1989, 394-401.
8. Nielsen, J. Assessing the learnability of HyperCard as a programming language. Manuscript submitted for publication 1990.
9. Nielsen, J. and Molich, R. Teaching user interface design based on usability engineering. *ACM SIGCHI Bulletin* **21**, 1 (July 1989), 45-48.
10. Smith, S.L. and Mosier, J.N. *Guidelines for Designing User Interface Software*. Report **MTR-10090**, The MITRE Corp, Bedford, MA, August 1986.

Extracting Usability and User Experience Information from Online User Reviews

Steffen Hedegaard
Department of Computer Science,
University of Copenhagen
Njalsgade 128,
2300 Copenhagen S, Denmark
steffenh@diku.dk

Jakob Grue Simonsen
Department of Computer Science,
University of Copenhagen
Njalsgade 128,
2300 Copenhagen S, Denmark
simonsen@diku.dk

ABSTRACT

Internet review sites allow consumers to write detailed reviews of products potentially containing information related to user experience (UX) and usability. Using 5198 sentences from 3492 online reviews of software and video games, we investigate the content of online reviews with the aims of (i) charting the distribution of information in reviews among different *dimensions* of usability and UX, and (ii) extracting an associated vocabulary for each dimension using techniques from natural language processing and machine learning. We (a) find that 13%–49% of sentences in our online reviews pool contain usability or UX information; (b) chart the distribution of four sets of dimensions of usability and UX across reviews from two product categories; (c) extract a catalogue of important word stems for a number of dimensions. Our results suggest that a greater understanding of users' preoccupation with different dimensions of usability and UX may be inferred from the large volume of self-reported experiences online, and that research focused on identifying pertinent dimensions of usability and UX may benefit further from empirical studies of user-generated experience reports.

Author Keywords

User experience; usability; natural language processing; end user reviews; machine learning.

ACM Classification Keywords

H.5.2 Information Interfaces and Presentation (e.g., HCI): User Interfaces–Evaluation/Methodology

General Terms

Experimentation; Human Factors; Measurement.

INTRODUCTION

Investigation of a product for usability or user experience (henceforth *UUX* for short) problems typically requires expensive experimentation. In contrast, informal, but structured electronic word-of-mouth communication between end users affords a potential, cheap source of information concerning UUX. Customer reviews on commercial sites such as `amazon.com`, or dedicated review sites such as `epinions.com`, contain reviews that are not just summary assessments or recommendations, but also self-reports of the end users experiences, in their own words, in the wild.

The aim of this paper is to quantify the amount of UUX information and dimensions in online reviews from the specific domains of software and video games. We also implement and test a machine-learning-based classifier that tags sentences in reviews according to whether they contain usability or UX-related information and according to the dimensions of usability or UX they pertain to. The primary aim of the classifier is to automatically extract the pertinent vocabulary of end users associated with the various dimensions of UUX. A secondary aim is to investigate the feasibility of using such a classifier to automatically catalogue UUX information found in databases of thousands of reviews, too large for qualified human analysis. We hope to aid the understanding of which dimensions of product use motivate laymen reviewers, and in the future potentially use this understanding when re-designing a product. The scope of the present work is to provide a tool to UUX researchers; future work will explore the automatic identification and extraction of specific actionable outcomes for practioners.

In order to process information from many different reviews, our approach focuses on extraction of information from individual sentences, rather than entire texts. This is somewhat at odds with approaches in focusing on obtaining a holistic understanding of interaction [16], but as a review may incorporate both good and bad experiences relating to many different dimensions of UUX, we believe that a sentence-based bottom-up approach will yield *more precise* information about the "typical" vocabulary associated to specific dimensions of UUX.

Related work

User experience has been studied by soliciting user narratives [15, 25, 32] where information is manually extracted from user-generated texts. The volume of texts studied has been substantial (500 texts in [15]), but still small enough for dedicated researchers to process manually, and the users have been specifically asked to write the texts, unlike the typ-

CHI 2013, April 27–May 2, 2013, Paris, France.

ical online review. Similarly, studies in asynchronous usability testing and reporting [7, 6] have studied user-generated problem reports that are only later reviewed by experts or researchers.

Outside UUX, classification and information extraction from user-generated text is a vibrant research area, both for full texts, and at sentence or short message level, e.g. tweets [39]. Some pertinent examples: Gamon et al. [13] perform sentence-level sentiment analysis on car reviews using several methods from machine learning, but only classify sentences into positive/negative/other. Kim et al. [24] perform sentence-based classification of *pros* and *cons* for mp3 and restaurant reviews in order to extract plausible reasons for the reviewers' recommendations or non-recommendations, but they do not extract vocabulary or classify according to a diverse number of dimensions. Pang et al. [33] classify movie reviews as positive/negative at document level. The primary differences between our work and studies on sentiment analysis as above are twofold: Firstly, we focus on a substantial number of distinct UUX dimensions, some of which may be objective (e.g. a count of false positives from AV software) with no negative or positive opinion, and others which may be described in neutral terms (e.g. aesthetics) where a reviewer can use neutral terms without showing any sentiment. Secondly, unlike most studies in sentiment analysis, the outcome of the classification is primarily a means to an end, namely charting the UUX content of reviews and the vocabulary used by reviewers to describe UUX-related phenomena.

REVIEWS ON THE INTERNET AND UUX

For the purpose of this paper a *review* is a piece of text detailing pros and cons of a product and possibly an assessment of it and recommendations for potential buyers, written by a user of the product who has been in possession of said product and used it for some time. It may be written either by a professional reviewer or an ordinary end user. We concentrate on reviews assumed to be written by ordinary end users on dedicated web sites, for example `epinions.com` or `amazon.com`.

An example of an online review is shown in Figure 1.

Consider the following sentence from a review of the game Gears of War for the Xbox 360:

"You'll be a little creeped out while playing this game almost all the time."

The above sentence clearly contains information that is *hedonic* in nature: Being scared due to the horror elements in the game, and there is a–much less clear–element of the *satisfaction* usability aspect: the sentence communicates a *positive* experience by the user.

From a communication perspective, user reviews may be viewed as *word of mouth* communication: informal communication between private parties concerning evaluation of goods and services [1]; reviews from review sites, online fora, and blogs are clearly examples of such informal communication, and are accordingly called eWoM (electronic word of mouth) in the literature [17]. Anderson [1] found that either

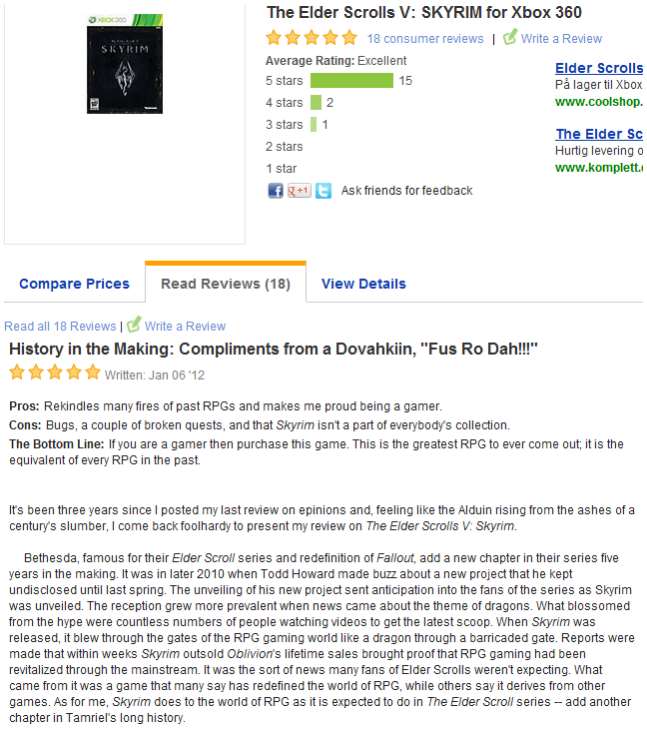


Figure 1. Example of the first part of a review from epinions.com 12. September 2012.

very satisfied or very dissatisfied customers were more likely to engage in (non-electronic) word of mouth and the word of mouth satisfaction was best described by a bimodal (U-shaped) model. This was confirmed for online reviews by Hu et al. [20] who found that 53% of products on Amazon had a bimodal score distribution (with peaks at very low scores and very high scores). Due to the bimodal distribution, the average score for these products may be misguiding. In addition, the information extracted from online reviews may not be indicative of the experience of the average user, but may rather represent those experiences that add or deduct so much from certain users' experience that they are motivated to write a review.

For (non-electronic) word of mouth communication, extremely dissatisfied customers also engage more in word of mouth than very satisfied customers, though in a sizeable case of their data the differences were not significant[1]; it seems plausible that the same phenomena occur for online user reviews. There is evidence that potential buyers put more emphasis on reviews with low satisfaction than those with high satisfaction as they have a bigger impact on product sales than that of positive reviews and word of mouth [8, 27].

Usability and UX in reviews

Usability is a way to measure a products ability to help a user solve a given task adequately. It is dependent on the product, task, user and circumstances [21], and has been the object of intense academic scrutiny. UX is a younger, emerging field that studies users' experience with products and the design of

such product with the purpose of generating certain experiences [15, 2].

In usability, traditional studies focus on short term product use (median 30 minutes duration [19]) and conducted in lab settings; few studies stretch across longer time periods and then only weeks [19]. In contrast, most UX research concerns open use situations (61%) and controlled task (33%) experiments, only 20% of papers contain studies based on user-initiated use [2]. No UX research covers longer time periods of months or years which is the expected life span of most products but instead covers at most only a few weeks [2].

In contrast to traditional studies, reviews describe a users opinion and experiences after more protracted use. And in contrast to user narratives solicited for product improvement or research purposes (cf. [15, 25, 32]) online reviews are in a different genre: Authors must follow certain conventions of the review genre, for instance give recommendations on whether or not to buy it. In addition, and unlike narratives written for UX studies, customer reviews on Internet sites appear to be written because the reviewer is motivated by his or her own use of the product, usually in conjunction with some small reward (tangible if the review site offers "credit" for reviews, intangible in the form of community recognition because of the perceived help afforded by a review, or both).

There are important caveats when assessing the potential usefulness of online reviews: It is not clear whether online reviews are written by users typical of the user base; in addition, very few details about reviewers (e.g. gender, age, preferences) are available, in contrast to standard usability studies. Furthermore, some reviews may be fake. Finally, the bimodal distribution of satisfaction present in word of mouth communication leads us to conjecture that in terms of satisfaction, the average user is underrepresented among reviewers, and that reviews may not always yield a representative description of the typical experiences among the user base. However, satisfaction extremes are well represented, it should thus be possible to extract information about situations where the product under review performs both bad and good.

DIMENSIONS OF USABILITY AND USER EXPERIENCE

Usability and User Experience are central terms in human-computer interaction. Their precise definition, and their subdivision into dimensions such as *Efficiency*, *Learnability*, *Hedonic quality*, and so forth is still debated [19], in the case of UX hotly so [16, 23, 4], and there seems to be no universal consensus about whether UX is an aspect of usability or vice versa.

We are particularly interested in the way researchers have subdivided UUX into various *dimensions* that pertain to specific aspects, viewpoints, or phenomena within UUX. We briefly review existing research below.

Dimensions of usability

The ISO 9241 standard [21] defines usability in terms of the three dimensions *effectiveness*, *efficiency* and *satisfaction*:

> Usability: The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.

A more fine-grained description of usability is obtained by the following five dimensions which, with some variation in the naming of the aspects, are often used for measuring and describing usability in models and literature [12, 36]: (i) *Effectiveness/Errors* [9, 38, 31, 34, 37, 21], (ii) *Efficiency* [9, 21, 38, 31, 34, 37], (iii) *Satisfaction* [9, 21, 38, 31, 34, 37], (iv) *Learnability* [9, 38, 31, 34, 37], (v) *Memorability* [9, 38, 31, 37].

The definitions of the above five dimensions vary somewhat in the literature, and some studies use only a subset of the above [19]. In addition, some studies use a precise and limited definition of measures, and others such as the ISO definition [21] take a broad view of the measures.

Dimensions of user experience

Unlike usability, there seems to be much less consensus on the definition of the notion of user experience and its segmentation into meaningful aspects.

The ISO 9241-210[22] definition states:

> User experience: A person's perceptions and responses that result from the use and/or anticipated use of a product, system or service.

We may interpret the above as being covered by the *Satisfaction* dimension of the ISO 9241-11 definition of usability [4], but the literature contains many more nuanced interpretations: For example, Bevan [3] describes four dimensions, called *satisfaction measures*, and Ketola and Roto [23] describe a study at Nokia among relevant senior staff who were asked which UX data they found useful, which Bevan later grouped into dimensions [4]. McNamara et al. [29] split the need for evaluation into the three (overlapping) components of *functionality*, *usability* and *experience*. Similarly, Hassenzahl [16] divides UX analysis into the three partially overlapping approaches *beyond the instrumental*, *emotion and affect*, and *the experiental*, but in later work [14] shifts focus to the subjective side of product use and relates it to Self Determination Theory [35] and Flow [11]. Bargas-Avila and Hornbæk [2] systematically collect a sample of 51 publications from 2005-2009 reporting empirical studies on UX and describe a number of (non-mutually exclusive) dimensions together with the percentage of studies from their sample that pertain to the identified dimensions.

Dimensions selected for this paper

Based on the literature above, we elected to use the 5 standard dimensions of usability, and chose the sets of dimensions from 3 of the studies of UX that had both precise definitions of the dimensions and clear demarcations of the differences between them (with two exceptions in FREQUENT, see below).

A summary is shown in Table 1; we briefly describe the dimensions below.

CLASSICUA ([12, 36])	BEVAN ([3])	KETOLA ([23, 4])	FREQUENT ([2])
Dimension	*Dimension*	*Dimension*	*Dimension*
Memorability	Likeability	Anticipation	Affect and Emotion
Learnability	Pleasure	Overall Usability	Enjoyment, Fun
Efficiency	Comfort	Hedonic	Aesthetics, Appeal
Errors/effectiveness	Trust	Detailed usability	Engagement, Flow
Satisfaction		User differences	Motivation
		Support	Enchantment
		Impact	Frustration
			Hedonic

Table 1. Dimensions of UUX used for the studies in this paper. . CLASSICUA is short for "classic usability", BEVAN and KETOLA are named after the authors of the pertinent studies, and FREQUENT is short for "frequently mentioned dimensions".

The dimensions of CLASSICUA [12, 36] are: *Errors/effectiveness*: The number of (non-fatal) errors made by users on their way to completing a task or the quality of task outcome. *Efficiency*: The speed or other measure of cost associated with performing the task for users at a given experience level. *Satisfaction*: A subjective rating of satisfaction with product use or liking of the product or features. *Learnability*: The amount of time it takes to learn to use the system, how difficult it is for a first time user or development over time. *Memorability*: How well users retain information gained about or through the system.

The dimensions of BEVAN [3] are: *Likability:* The extent to which the users are satisfied with their perceived achievement of pragmatic goals, including acceptable perceived results of use and consequences of use (note the close similarity with *Satisfaction* from CLASSICUA). *Pleasure:* The extent to which the users are satisfied with their perceived achievement of hedonic goals of stimulation, identification, evocation and associated emotional responses. *Comfort:* The extent to which the users are satisfied with physical comfort. *Trust:* the extent to which the users are satisfied that the product will behave as intended.

The dimensions of KETOLA [23, 4] are: *Anticipation*: What did users expect, what is the anticipated use? *Overall usability*: Was the user successful in taking the product into use or upgrading from a previous product. *Hedonic*: Fulfillment of inner needs such as pleasure, enjoyment, or things preventing this such as frustration. *Detailed usability*: Going into details on which functions are used, ordinary usability problems and performance satisfaction/problems. *User differences*: Differences between users such as previous product experience, how they access features and differences between the actual buyers and target user group. *Support*: Aspects with the human- or software-support available and how it affects user satisfaction, possible product returns, or user wish lists. *Impact of use*: If and how the new device change the usage patterns of the users.

The dimensions of FREQUENT are: *Affect and Emotion*: Affect and emotion induced by using the product, including other aspects such as Enjoyment, fun and Frustration. This dimension fully encompasses *Enjoyment, fun* and *Frustration*, and would be considered encompassed by the *Hedonic* dimension. *Enjoyment, Fun*: How entertained is the user while using the product? This is also an affect and emotion, and hedonic, dimension. *Aesthetics, Appeal*: Appreciation of beauty or good taste. Typically associated with graphics or sound. *Engagement, Flow*: How engaged is the user in using the product forgetting everything else? Also includes challenge versus skill balancing needed for achieving flow state. *Motivation*: What motivates the user in using the product (task/inner motivation etc.)? *Enchantment*: Being "both caught up and carried away" in the experience forgetting everything else, and causing a disorientation associated with a pleasurable sense of fullness and liveliness that charges attention and concentration. *Frustration*: Frustration or hardship induced by using the product. This is also a negative hedonic dimension. *Hedonic*: Defined the same way as in KETOLA. We discarded the two dimensions *Generic UX* and *Other* as reported by [2] as no clear definition was available.

PRE-STUDY

To see whether randomly sampled reviews contain sufficient UUX-related information to warrant further study, we performed a pre-study among usability experts who were given a sample of Internet reviews and a free-form exercise asking them to mark sentences containing information about usability or UX.

Participants

9 usability experts were contacted, all of whom are active researchers in usability (7 from academia, 2 from industry). Of these, 8 gave affirmative answers and were enrolled as participants. All participants were compensated with two bottles of wine.

Procedure

24 reviews were sampled on January 5 2012 from 12:00 – 14:00 from the website `epinions.com`, collecting the 6 most recent reviews of each of the four categories "Digital Cameras", "Headphones", "Software" and "Video games". 3 reviews were discarded, all from the software category (1 was a review of an iPhone game (games were covered explicitly in another category), and 2 were reviews of printed children's books). Each review was randomly assigned to 2 distinct participants.

Each participant was asked to read and comment on six different reviews in total. The participants received only written instructions asking them to free-form annotate text in the reviews that they found interesting concerning (a) their own perception of usability, and (b) user experience. Participants were neither given definitions of usability or user experience, but encouraged to use their own perception of these terms. Each participant was asked to use at most 2 hours in total on all six reviews, including time to read and to annotate.

We collected the annotated texts and post-processed them in two ways:

Raw containment of UUX: Each review was manually split into sentences and was marked with the identity of a participant if the participant had marked part of or the entire sentence as relevant. Due to the level of annotation performed by most experts, no distinction was done between dimensions or UX and usability based on the experts comments.

Presence of UUX dimension: For each review, the first author coded all sentences annotated by at least one participant using the dimensions from Table 1; the same sentence could be annotated with more than one dimension. Examples of dimension assignments can be seen in Table 2

Content	Dimensions present
If you like multiplayer strategy games, buy this with confidence.	satisfaction, user differences
Those expectations were met. Mostly, anyway.	anticipation, satisfaction
... making the game enjoyable for beginners as well as veterans.	user differences, flow, enjoyment, hedonic
Multiplayer is excellent, but the single player campaign isn't.	satisfaction
Most of the inter-mission story telling happen in this mode, which tend to be awkward and clumsy.	satisfaction, frustration
Most of the missions are enjoyable, and each one has optional goals which add replay value.	enjoyment, hedonic, engagement/flow

Table 2. Examples of annotation from the video game Starcraft 2.

Results

Raw containment of UUX: Calculation of inter-rater agreement for raw containment of usability or UX indicated that participants were somewhat in agreement on which sentences did, or did not, contain any relevant information at all, but that no hard conclusions should be drawn based on the data (Krippendorff's $\alpha = .783$)[1].

In total, 13 % of all sentences were marked as relevant to usability or UX by both participants assigned to each review, 36 % as relevant by one, but not both assigned to each review, and 51% of all sentences were unmarked (i.e., deemed as irrelevant by participants).

There was great diversity in the understanding of UUX and annotation volume per participant. One participant specifically noted that he had given up marking user experience data as it "virtually encompassed everything", and only a single participant consistently annotated UX and usability information as two distinct categories. We observed some discrepancies in annotations; for example, one participant had marked the sentence "The product works extremely well" as relevant for UUX in a review, but later in the review failed to mark the similar sentence "It also works well when listening to music while using power tools . . ." as relevant.

Presence of usability or UX dimension: The results are summarized in table 3.

For the classic usability measure as seen in Table 3, almost all sentences in the dimension *errors/effectiveness* were describing quality of task outcome (e.g., music quality for headphones), but a few classic error counts were also present (e.g., notes correctly transcribed by a sheet music scanning feature of a program, and false positives in anti virus software).

Only rarely ($N = 4$) did reviews attach any numbers to measures of efficiency and effectiveness, and even then they were not considered as exact measures, but merely rough estimates such as "...and the whole process only takes about 5 minutes..."

Detailed inspection of the reviews revealed that some dimensions only occurred in specific product categories: The dimension *physical comfort* was exclusively encountered in camera and headphones reviews, and the dimension *pleasure* mainly for video games.

The dimensions of *motivation* and *enchantment*, both popular dimensions in empirical user experience research being represented in 8% and 6% of papers respectively [2], were not encountered at all in the pre-study.

CLASSICUA		BEVAN		KETOLA		FREQUENT	
Dimension	*Occ.*	*Dimension*	*Occ.*	*Dimension*	*Occ.*	*Dimension*	*Occ.*
Memorability	0.04%	Likeability	4.70%	Anticipation	3.57%	Affect and Emotion	0.24%
Learnability	3.77%	Pleasure	0.56%	Overall usability	0.16%	Enjoyment, Fun	0.40%
Efficiency	1.44%	Comfort	1.04%	Hedonic	1.65%	Aesthetics, Appeal	0.48%
Errors/Effectiveness	3.61%	Trust	1.12%	Detailed usability	21.07%	Engagement, Flow	0.68%
Satisfaction	4.70%			User differences	2.17%	Motivation	0.00%
				Support	0.64%	Enchantment	0.00%
				Impact	0.32%	Frustration	0.20%
						Hedonic	1.65%

Table 3. Occurrences of dimensions found in sentences annotated by participants as a percentage of the total number of sentences in all reviews.

In summary, 13% + 36% = 49% of all sentences were marked as relevant by at least one of the two participants annotating each review. Some confirmation bias may be present as participants were specifically asked to look for information relevant to usability or user experience, but based on the results we concluded that the *volume* of text in a review relevant to usability or UX, and the *dispersion* of text across UUX dimensions were both substantial enough to warrant a larger-scale annotation experiment.

FIRST STUDY: ANNOTATION OF REVIEWS

Based on the promising results of the pre-study, we opted to harvest a larger sample of reviews and annotate them. We decided to keep the per-sentence annotation of the pre-study and concentrate on only 2 product categories as it would allow us to annotate more sentences in each product category while retaining the ability to make comparisons across categories.

Procedure

We collected reviews from the two product categories *Software* (520 reviews) and *Video games* (2972 reviews across various PC and console platforms) on the `epinions.com` website on July 5th, 2012. All public available reviews in the two categories were collected. We split each review into sentences using a routine from the Python NLTK [5] which came pre-trained on the British National Corpus. We then drew sentences randomly from the pool of all sentences above and performed manual annotation on each drawn sentence. In total 4587 sentences (of a pool of 132609) were annotated from the *Video games* and 611 (of a pool of 18646) from the *Software*.

	Min.	Max.	Median	*M*	*SD*
Software:	26	11686	471	780.6	952.2
Video Games:	30	23989	721	880.8	774.4

Table 4. Word count statistics for reviews.

The annotation was conducted by one of the authors and two graduate student annotators: The graduate students were given a short, written introduction description of all dimensions and participated in a co-annotation workshop for four

[1]No hard conclusions should be based on data with $.667 \leq \alpha < .8$ [26]

hours with one of the authors acting as instructor and senior annotator. This was followed by four hours of individual annotation where the student annotators could freely consult the senior annotator for questions. Inter-rater agreement was computed at the end of the individual annotation with Krippendorff's α for all dimensions in the $[0.9 - 1.0]$ range. Each graduate student annotator then continued individually for 22-25 hours over the course of two weeks, and the senior annotator for 10 hours during one week.

All annotations were performed in a custom-built tool built by a research programmer not otherwise involved in the study.

A total of 6655 sentences were annotated with 1315 sentences annotated by the senior annotator 2922 sentences by annotator 2, and 2418 sentences annotated by annotator 1, every tenth sentence annotated by an annotator was randomly chosen from sentences already annotated by another in order to compute inter-rater agreement; all dimensions annotated showed a high degree of agreement (Krippendorff's α lowest score $\alpha = .842$; 22 of 25 dimensions had $\alpha > .90$)[2].

Some examples of sentences from reviews and their annotations:

- *"Once again, the way sound distorts during the slow-motion sequences adds a nice touch to the experience."*. This sentence mainly describes *Aesthetics* and subsequently the effect on *Engagement* while playing the game, yet also expresses *Satisfaction* with the effect. As this example illustrates, several dimensions are often encountered together with *Satisfaction*.
- *"The sound is what you'd expect from a Nintendo title, and can become quite annoying after extended plays."* This sentence describes displeasure, relevant to the dimension *Pleasure*; this is also a measure of *Satisfaction* with the sound.

Results

Table 5 shows the fraction of sentences in each product category annotated by the various UUX dimensions.

The table confirms the observation from the pre-study that some dimensions are hardly used at all: In CLASSICUA, *Memorability* does not occur at all in the software category, and only in 0.37% of sentences among video games. Likewise, *Efficiency* occurs very rarely (4.45% in the software category, 1.12% in video games). Among the dimensions from BEVAN, *Comfort* and *Trust* are again absent, but *Likeability* is prominent, as expected from the pre-study. The dimension *Pleasure* is more prevalent among video game reviews than reviews of software.

In KETOLA, *Impact* is almost completely absent, and *Support* rare, but more prevalent among software products, possibly reflecting that this dimension is more valued among users of software products than video games. Conversely, the dimension *Hedonic* is present in 7.77% of all sentences sampled in the video games category, more than twice as often as in the software category.

[2]Data with $\alpha > .8$ are generally considered reliable [26]

It is striking that there are quite few sentences annotated by dimensions in the FREQUENT classification (9.25% for software, 29.72% for video games) compared to the CLASSICUA, BEVAN and KETOLA categories (where more than 40% of all sentences contain usability or UX information). This phenomenon is due to the FREQUENT classification's lack of a "catch-all" category such as CLASSICUA's *Satisfaction*, BEVAN's *Likability* and KETOLA's *Detailed usability*.
In summary, the results support two clear conclusions: First, The product domain of the review (in our case, software, resp. video games) influences the amount of *sentences* that pertain to specific dimensions of UUX (e.g., the dimension *engagement, flow* is much more prevalent in reviews of video games than in reviews of other software). Second, the four sets of dimensions of UUX we consider differ very much in the balance of dimensions: Clearly, FREQUENT, the model containing the most categories, has the most even distribution of sentences across the various dimensions, whereas the other sets of dimensions seem to have greatly skewed distributions towards the "catch-all" categories described above.

SECOND STUDY: AUTOMATIC CLASSIFICATION OF SENTENCES IN UUX DIMENSIONS

Based on the results of the first study, we wish to investigate the vocabulary employed by users when conveying information relevant to the dimensions of UUX. One straightforward way of extracting such a vocabulary is to construct a machine learning classifier that discriminates between dimensions based on words or other features of the text that are automatically computed during the training of the classifier. An added benefit is that the constructed classifier, if precise, may be used for automatically tagging a sentence with the UUX dimensions it pertains to. This tagging task may be viewed as a set of binary classification tasks: For each dimension, and for each sentence, does the sentence pertain to that dimension, or not. Such a tagger may be used to either aid future researchers in manual annotation, or in *lessening* the amount of sentences to be studied (i.e., sentences receiving no tags by an automatic classifier can be ignored at little risk).

Procedure

For each UUX dimension a binary classifier using a bag-of-words feature set was trained and evaluated in a sequence of steps as follows: *Preprocessing step:* Each sentence was tokenized, words from the NLTK stop word list [5] removed and the remaining words stemmed using the `Snowball` stemmer. *Data split step:* The total dataset is split into a five-fold stratified cross-validation [28][3] scheme. *Training step:* For each cross-validation split, the training set is used for creating feature vectors with TFxIDF weighting and χ^2 [28] ranking is subsequently used to discard the 10% worst discriminating features. A linear kernel Support Vector Machine (SVM) [10] is then trained using the feature set. *Classification and validation step:* For each cross-validation split, classification of the evaluation set using each SVM classifier is performed, and the classification results for all the cross-validation splits are aggregated, and standard performance measures (see Table

[3]Stratified cross validation maintains the same class balance in the training and evaluation sets as found in the total data set.

CLASSICUA			BEVAN			KETOLA			FREQUENT		
Dimension	*Software*	*Video Games*	*Dimension*	*Software*	*Video Games*	*Dimension*	*Software*	*Video Games*	*Dimension*	*Software*	*Video Games*
Memorability	0.00%	0.37%	Likability	**31.34%**	**34.57%**	Anticipation	2.23%	3.78%	Affect and Emotion	**2.91%**	**8.63%**
Learnability	**7.36%**	**3.54%**	Pleasure	**2.23%**	**6.08%**	Overall Usability	1.37%	0.71%	Enjoyment, Fun	**1.37%**	**6.81%**
Efficiency	**4.45%**	**1.12%**	Comfort	0.00%	0.17%	Hedonic	**3.42%**	**7.77%**	Aesthetics, Appeal	**3.60%**	**11.70%**
Errors/effectiveness	**17.64%**	**8.61%**	Trust	0.00%	0.17%	Detailed usability	44.52%	41.34%	Engagement, Flow	**1.71%**	**12.24%**
Satisfaction	**31.34%**	**34.57%**	*Any Dimension*	***32.02%***	***37.23%***	User differences	**12.33%**	**8.61%**	Motivation	0.86%	1.42%
Any Dimension	***46.58%***	***41.83%***				Support	**2.74%**	**0.52%**	Enchantment	**0.00%**	**0.86%**
						Impact	0.00%	0.26%	Frustration	1.20%	1.37%
						Any Dimension	*53.60%*	*50.83%*	Hedonic	**3.42%**	**7.77%**
									Any Dimension	***9.25%***	***29.72%***

Table 5. Distribution of dimensions in sentences within Software and Video Games reviews. The "Any dimension" rows indicate the percentage of sentences annotated with at least one dimension. Each sentence can be annotated with more than one dimension, hence "Any dimension" is not the sum of the other numbers in the same column. Differences between the *Software* and *Video Games* categories were tested for significance using the non-parametric two-tailed Wilcoxon rank-sum test [18] and significance at $p < .05$ is indicated in boldface.

6) calculated. *Extraction of important words for each dimension:* For each dimension, we extracted the most informative words by selecting the word stems having the largest distance in descending order to the separating hyperplane afforded by the SVM.

Aside from the classic *bag of words* approach as described above, we also experimented with the following feature sets commonly used in text classification tasks: binary bag of words, word di-grams and tri-grams, a combination of tri-agrams and a feature set consisting of all possible `Wordnet` synsets [30], and `Wordnet` synsets with automatic part of speech (POS) tagging. All of the alternative feature sets had slightly worse average performance than an ordinary bag of words approach, hence were discarded.

To avoid drawing erroneous conclusions from unreliable data, we elected to only consider the extracted word stems from the dimensions where the classifier performed better than random chance. This was tested against a baseline classifier that always assigns to the majority class with significance at $p < .05$ (using the non-parametric McNemar's test with Yates' correction, as we have categorical data and cannot assume a specific prior distribution).

Results: Quality of the classifier

To evaluate the quality of the classifier, we use the classic information retrieval metrics *precision*, *recall*, and *F1* [28].

Precision is the fraction of sentences correctly classified as relevant for a dimension among all sentences classified as relevant for it; *recall* is the fraction of sentences actually relevant for a dimension that are also correctly classified as relevant for it; *F1* is the harmonic mean of precision and recall (see Table 6). The results for the various dimensions are shown in

Precision	Recall	F1
$P = \frac{\lvert R_d \cap C_d \rvert}{\lvert C_d \rvert}$	$R = \frac{\lvert R_d \cap C_d \rvert}{\lvert R_d \rvert}$	$2\frac{P \cdot R}{P+R}$

Table 6. Definitions of precision, recall and F1. R_d is the set of sentences relevant for dimension d, C_d is the set of sentences that the classifier tags as relevant for d.

Table 7, Significant results are marked in bold.

Table 7 shows that for the dimensions that are barely represented in the data, the classifier *all* sentences are classified as *not* relevant for the dimension. This is the case for *Memorability, Efficiency, Comfort, Trust, Overall usability, Support, Impact, Motivation and Enchantment* that all have precision and recall at zero flat. For the more commonly occurring dimensions, the classifier performs better than the baseline of assigning all sentences to the majority class, but it is clearly quite conservative: Precision values are generally high, but recall values low (e.g., for *Learnability*, *Anticipation*, *User differences*, *Engagement, Flow* and *Hedonic*). In short, for these dimensions, the sentences tagged as being relevant to a dimension will be relevant with high probability, but the classifier will miss many relevant sentences. For dimensions that commonly occur in the data, the classifier works well, as should be expected: precision, recall and F1 are all high for *Satisfaction* and *Likeability* and–again with the exception of recall–for *Enjoyment, Fun* and *Affect and Emotion, Frustration*, and *Hedonic*.

Differences in both data domain and classification tasks preclude us from directly comparing to other studies, but for all but the very sparsely represented dimensions, the performance of the classifier is on par with studies conducting sentence-based classification: Gamon et al. [13] performed sentence level sentiment analysis on car reviews with precision for the negative class from 0.85 to 0.55 with recall from 0.1 to 0.25. Similarly, Kim et al. [24] classified sentences with regard to *pros* and *cons* content, achieving $P = 0.59, R = 0.62, F1 = 0.61$ and $P = 0.54, R = 0.52, F1 = 0.53$ respectively on a well-balanced dataset of hotel reviews.

Results: Vocabulary of reviews

Table 8 holds the 30 most important word stems for each dimension where the difference in precision, recall and $F1$ between the classifier and the baseline was significant. As an example of top word stems associated with a dimension *not* included in the table are "sooth", "cute", "reliev", "handhold" and "exist", all associated with the dimension *trust*.

For the dimension *Frustration* with word stems "frustrat", "incompatibilit", "hardest", "perpetu", "insult" what seems like less relevant words also made it to the top 30, for instance "babysit". This is due to reviews containing text such as *"And that is of course an AI partner controlled friend, there is nothing that can ruin a good RPG then a partner that is supposed to be helping you but instead makes you feel like your babysitting a 5 year old with mental problems, on top of battling blood thirsty monsters."* This particular sentence also illustrates the intricacies of our task: Clearly, the use of "babysit" is figurative, not literal, hence signals frustration.

CLASSICUA				BEVAN				KETOLA				FREQUENT			
Dimension	*P*	*R*	*F1*	*Dimension*	*P*	*R*	*F1*	*Dimension*	*P*	*R*	*F1*	*Dimension*	*P*	*R*	*F1*
Memorability	0.00	0.00	0.00	Likeability	**0.69**	**0.47**	**0.56**	Anticipation	0.61	0.06	0.11	Affect and Emotion	**0.81**	**0.42**	**0.56**
Learnability	**0.93**	**0.08**	**0.15**	Pleasure	**0.77**	**0.44**	**0.56**	Overall Usability	0.00	0.00	0.00	Enjoyment, Fun	**0.92**	**0.58**	**0.71**
Efficiency	0.00	0.00	0.00	Comfort	0.00	0.00	0.00	Hedonic	**0.80**	**0.41**	**0.54**	Aesthetics, Appeal	**0.82**	**0.47**	**0.59**
Errors/effectiveness	**0.73**	**0.05**	**0.09**	Trust	0.00	0.00	0.00	Detailed usability	**0.67**	**0.54**	**0.60**	Engagement, Flow	**0.72**	**0.16**	**0.26**
Satisfaction	**0.69**	**0.47**	**0.56**	*Any Dimension*	***0.73***	***0.70***	***0.71***	User differences	0.68	0.04	0.07	Motivation	0.00	0.00	0.00
Any Dimension	***0.80***	***0.51***	***0.62***					Support	0.00	0.00	0.00	Enchantment	0.00	0.00	0.00
								Impact	0.00	0.00	0.00	Frustration	**1.00**	**0.36**	**0.53**
								Any Dimension	***0.72***	***0.54***	***0.62***	Hedonic	**0.80**	**0.41**	**0.54**
												Any Dimension	***0.80***	***0.51***	***0.62***

Table 7. Automatic classification of dimensions. Results are checked for significance against a baseline classifier that assigns to the majority class. Significance is calculated using the non-parametric McNemar's test with Yates' correction for continuity [18]. Significant results ($p < .05$) are marked in boldface.

Other spurious word stems in Table 8 (e.g., "sc", "thus", "gps") can be attributed to two phenomena: (i) the word stems in the dimension have low discriminatory power, whence the classifier was barely able to distinguish relevant/irrelevant sentences, and (ii) a word stem may be considered important if it by chance occurs in the training corpus of the classifier in a small number of sentences, all of which are relevant to the dimension.

Dimensions such as *Hedonic*, *Pleasure* and *Affect and emotion* which fully or partially encompas other dimensions tend to share most of the top places of the included subcategories. For example, nine of the ten most important word stems from the dimension *Enjoyment and Fun* were found among the top 15 important word stems in the encompassing dimension *Effect and Emotion* which also rated other word stems such as "scari" and "frustrate" related to emotion, but *not* enjoyment highly. The notable exception to this phenomenon is the dimension *Detailed usability* that encompasses all measures of classic usability as well as mention of specific usability problems; this dimension has many highly ranked word stems relating to *Satisfaction*, but fewer highly ranked words stems also occurring among the other classic usability dimensions.

DISCUSSION

We have unearthed a significant difference between software and video game reviews in terms of which UUX dimension they frequently mention. With the exception of *satisfaction* software reviews emphasize classic usability measures more than video game reviews, which in term put more emphasis on dimensions such as *Hedonic, Affect and emotion, Pleasure* and *Enjoyment and fun*. The notable exception to this difference is *Frustration* for which the difference between software and video games were not significant.

The automatic classifier works well on commonly occurring dimensions, but tends to be too conservative even for these dimensions. There are ways of improving such classifiers, but our experiments suggest that simple off-the-shelf machine learning solutions are insufficient. When sifting through large amounts of material, it is easy to miss infrequent, but potentially important information, for instance the presence of information pertaining to *Enchantment* or *Trust*, and the classifier clearly is unable to assist a human expert in this regard. However, for some of the commonly occurring categories, the classifier may assist by *removing* sentences irrelevant for the dimensions, at the cost of some potentially relevant sentences being removed.

The set of word stems extracted shows that some dimensions (e.g., *Enjoyment, Fun*) have associated vocabularies containing words closely related to the words used to *describe* the dimensions in the literature (e.g., word stems that are synonymous or antonymous to enjoyment and fun, describing respectively positive and negative experiences), but other dimensions such as *Errors/effectiveness* instead have vocabularies related to specific problems and errors such as "lag", "glitch", "imprecise" and "bug". This illustrates the varied vocabulary used by reviewers when describing specific dimensions of interaction, and that the vocabulary is more varied for some dimensions than others.
Our results strongly suggest that more complex information about how users *express* their feelings and experiences with situations and problems related to UUX can be extracted from reviews and other narratives. The results also suggest that the task of mapping the users' utterances to specific dimensions of UUX is only partially possible to do in an automatic fashion and that some of these dimensions are associated to a complex vocabulary.

When using sets of dimensions of UUX for *practical* purposes–for example for gauging which dimensions of a product are perceived to be important by users–then sets with many complementary dimensions such as FREQUENT appear to be more fine-grained and informative than other such sets. It may be possible for the UUX community to settle on a small, possibly domain-dependent set of dimensions, simply by performing empirical investigations such as ours, or in more traditional settings such as usability tests.

Limitations

While our results shed light on the general UUX concerns of end users, the anatomy of the reviews we considered, possibly Internet reviews in general, does not seem to contain much detailed information about specific situations of use, or of measurements. No reviewer writes "The number of mouse click to navigate from the start screen to the functionality I want is 7, and that this is annoying". While it is conceivable that a professional software reviewer, or an end user taking a conscious interest in usability, would write such a sentence, we did not encounter these. Thus, it seems highly unlikely that mining Internet reviews can supplant traditional usability testing or UX studies.

Learnability	Errors/effectiveness	Satisfaction	Hedonic	Detailed usability	Pleasure	Affect and emotion	Enjoyment, Fun	Aesthetics, Appeal	Engagement, Flow	Frustration
intuit	glitch	great	fun	realli	fun	fun	fun	graphic	challeng	frustrat
easier	issu	love	enjoy	great	enjoy	enjoy	enjoy	sound	addict	incompatibilit
learn	lag	realli	frustrat	best	bore	bore	bore	music	replay	hardest
figur	camera	worth	annoy	nice	love	frustrat	entertain	realist	difficulti	perpetu
easiest	imprecis	nice	bore	worth	entertain	annoy	funni	voic	hour	insult
eas	bump	best	funni	problem	annoy	entertain	love	soundtrack	difficult	injuri
straightforward	bug	graphic	love	love	felt	love	amus	effect	depth	dissadvantag
easi	configur	sound	entertain	overal	sooth	amus	humor	beauti	harder	nerv
sc	suspect	fun	humor	graphic	lighter	laugh	laugh	anim	moment	fuel
foreword	error	overal	hate	easi	workout	scari	excit	look	complex	afterward
practic	flaw	problem	sooth	recommend	grin	excit	sooth	environ	nonstop	grin
simpl	crowd	recommend	felt	issu	frustrat	addict	grin	vivid	tough	needless
angl	dodg	disappoint	hum	pretti	humor	humor	lighter	audio	easi	la
menus	biowar	definit	lighter	bad	chore	sooth	kinda	atmospher	valu	unfair
steam	ai	favorit	cute	disappoint	incompatibilit	felt	fell	color	deep	plain
sacr	troubl	good	grin	fun	fell	fell	hilari	visual	therefor	flat
plasmid	semblanc	bad	excit	sound	rooftop	grin	workout	sceneri	keep	grow
experiment	mater	price	catchi	good	nevertheless	lighter	nevertheless	sprite	imposs	fusion
password	inconsist	interest	addict	learn	laugh	chore	zero	impress	engag	cheat
straight	suffer	cool	workout	favorit	afterward	truli	intrigu	render	tire	melodramat
thus	data	fan	chore	definit	told	engag	rooftop	model	hard	habit
minut	confus	improv	tens	better	intrigu	tens	shatter	cute	intens	gasp
master	technic	fantast	incompatibilit	feel	nostalgia	kinda	told	detail	sc	heck
smooth	prompt	better	fell	price	shatter	incompatibilit	scare	appeal	painstak	annoy
nunchuk	resolut	perfect	afterward	improv	regardless	hilari	younger	realism	lenient	babysit
incred	respons	unfortun	nostalgia	interest	perpetu	workout	im	pixel	hardest	insan
sensor	load	lack	nevertheless	cool	moneybag	creatur	cute	bright	gripper	vito
gps	primit	pretti	stagger	would	everytim	cute	queue	stun	becom	slog
casual	precis	feel	intrigu	perfect	adict	nostalgia	tedious	creepi	interest	scaletta
applet	delay	qualiti	regret	qualiti	countless	nevertheless	jeremi	hear	most	flavor

Table 8. Informative word stems for each dimension in the video games and software corpora. Most informative word stems are at the top. Only dimensions for which the classifier achieved significant results are listed.

Finally, the sentence-based annotation has acted as a convenient proxy: If a user spends 10% of the sentences in a review discussing matters related to *Enchantment*, it is likely evident that enchantment is a major part of his view of the product; but there may be other measures that more precisely reflect *how much* the user is occupied with different dimensions of UUX.

Future work

The data we considered were limited to two specific domains (software and video games), and the volume of data, while respectable, was insufficient to establish a vocabulary for all usability dimensions. Future studies must extend our work to more domains, and must consider a very large volume of data (a rough estimate based on our work: several tens of thousands of sentences). In addition, the idea of extracting vocabularies and associating features of texts or other utterances to dimensions of UUX can be applied to other domains, including spoken words at traditional lab-based usability studies.

It seems worth to investigate the difference across product domains of the distribution of sentences among UUX dimensions found in this study. Likewise, it is interesting to link the presence of sentences pertaining to the UUX dimensions to attributes of the reviews that can be inferred otherwise, for example negative vs. positive reviews, or the helpfulness of reviews as voted upon by other users.

Filtering and grouping of dimensions may be examined in greater detail in follow-up studies also investigating actionable outcomes. Finally, a better-performing classifier, or human annotation of a larger amount of complete reviews instead of isolated sentences may allow for analyzing the distribution of the dimensions we consider, on a per-review basis.

REFERENCES

1. Anderson, E. Customer satisfaction and word of mouth. *Journal of Service Research 1*, 1 (1998), 5–17.
2. Bargas-Avila, J. A., and Hornbæk, K. Old wine in new bottles or novel challenges: a critical analysis of empirical studies of user experience. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, CHI '11, The ACM Press (2011), 2689–2698.
3. Bevan, N. Classifying and selecting ux and usability measures. In *International Workshop on Meaningful Measures: Valid Useful User Experience Measurement* (2008), 13–18.
4. Bevan, N. What is the difference between the purpose of usability and user experience evaluation methods. In *Proceedings of the Workshop UXEM'09 (Interact 09)* (2009).
5. Bird, S., Klein, E., and Loper, E. *Natural Language Processing with Python*. O'Reilly Media, 2009.
6. Bruun, A., and Stage, J. The effect of task assignments and instruction types on remote asynchronous usability testing. In *CHI*, J. A. Konstan, E. H. Chi, and K. Höök, Eds., ACM (2012), 2117–2126.
7. Castillo, J., Hartson, H., and Hix, D. Remote usability evaluation: Can users report their own critical incidents. In *Proceedings of CHI '98*, The ACM Press (1998), 253–254.
8. Chevalier, J., and Mayzlin, D. The effect of word of mouth on sales: Online book reviews. Tech. rep., National Bureau of Economic Research, 2003.
9. Constantine, L. L., and Lockwood, L. A. D. *Software for use: a practical guide to the models and methods of usage-centered design*. ACM Press/Addison-Wesley Publishing Co., 1999.
10. Cortes, C., and Vapnik, V. Support-vector networks. *Machine Learning 20* (1995), 273–297.

11. Csikszentmihalyi, M. *Flow: The psychology of optimal experience*. Harper Perennial, 1991.
12. Folmer, E., Van Gurp, J., and Bosch, J. A framework for capturing the relationship between usability and software architecture. *Software Process: Improvement and Practice 8*, 2 (2003), 67–87.
13. Gamon, M., Aue, A., Corston-Oliver, S., and Ringger, E. Pulse: Mining customer opinions from free text. In *Advances in Intelligent Data Analysis VI*, vol. 3646 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2005, 121–132.
14. Hassenzahl, M. User experience (ux): towards an experiential perspective on product quality. In *Proceedings of the 20th International Conference of the Association Francophone d'Interaction Homme-Machine*, IHM '08, The ACM Press (New York, NY, USA, 2008), 11–15.
15. Hassenzahl, M., Diefenbach, S., and Göritz, A. Needs, affect, and interactive products. *Human-Computer Interaction 25*, 3 (2010), 235–260.
16. Hassenzahl, M., and Tractinsky, N. User experience–a research agenda. *Behaviour & Information Technology 25*, 2 (2006), 91–97.
17. Hennig-Thurau, T., Gwinner, K., Walsh, G., and Gremler, D. Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the internet? *Journal of interactive marketing 18*, 1 (2004), 38–52.
18. Hollander, M., and Wolfe, D. *Nonparametric Statistical Methods*, 2nd ed. Wiley, 1999.
19. Hornbæk, K. Current practice in measuring usability: Challenges to usability studies and research. *Int. J. Hum.-Comput. Stud. 64* (February 2006), 79–102.
20. Hu, N., Pavlou, P. A., and Zhang, J. Can online reviews reveal a product's true quality? empirical findings and analytical modeling of online word-of-mouth communication. In *Proceedings of the 7th ACM conference on Electronic commerce*, EC '06, The ACM Press (2006), 324–330.
21. ISO 9241-11. Ergonomic requirements for office work with visual display terminals (vdts)–part 11: Guidance on usability. International Organization for Standardization, 1998.
22. ISO 9241-210. Human-centred design process for interactive systems. International Standards Organisation, 2010.
23. Ketola, P., and Roto, V. Exploring user experience measurement needs. In *5th COST294-MAUSE Open Workshop on Valid Useful User Experience Measurement* (2008).
24. Kim, S.-M., and Hovy, E. Automatic identification of pro and con reasons in online reviews. In *Proceedings of the COLING/ACL on Main conference poster sessions*, COLING-ACL '06, Association for Computational Linguistics (2006), 483–490.
25. Korhonen, H., Arrasvuori, J., and Väänänen-Vainio-Mattila, K. Let users tell the story. In *Proceedings of AMC CHI 2010 Extended Abstracts*, The ACM Press (2010), 4051–4056.
26. Krippendorff, K. *Content analysis: an introduction to its methodology*. Sage, 2004.
27. Lutz, R. Changing brand attitudes through modification of cognitive structure. *Journal of Consumer Research* (1975), 49–59.
28. Manning, C. D., Raghavan, P., and Schütze, H. *Introduction to Information Retrieval*. Cambridge University Press, 2008.
29. McNamara, N., and Kirakowski, J. Functionality, usability, and user experience: three areas of concern. *Interactions 13*, 6 (2006), 26–28.
30. Miller, G. Wordnet: a lexical database for English. *Communications of the ACM 38*, 11 (1995), 39–41.
31. Nielsen, J. *Usability Engineering*. Academic Press Inc, 1993.
32. Olsson, T., and Salo, M. Narratives of satisfying and unsatisfying experiences of current mobile augmented reality applications. In *Proceedings of ACM CHI 2012*, The ACM Press (2012), 2779–2788.
33. Pang, B., Lee, L., and Vaithyanathan, S. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02, Association for Computational Linguistics (2002), 79–86.
34. Preece, J., Rogers, Y., Sharp, H., and Carey, T. *Human Computer Interaction*, 1st ed. Addison-Wesley, Wokinham, England, 1994.
35. Ryan, R., and Deci, E. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American psychologist 55*, 1 (2000), 68.
36. Seffah, A., Donyaee, M., Kline, R., and Padda, H. Usability measurement and metrics: A consolidated model. *Software Quality Journal 14* (2006), 159–178.
37. Shackel, B. *Usability-context, framework, definition, design and evaluation*. Cambridge University Press, 1991, 21–37.
38. Shneiderman, B., Plaisant, C., Cohen, M., and Jacobs, S. *Designing the User Interface: Strategies for Effective Human-Computer Interaction*, 3rd ed. Addison-Wesley Publishing Company, 1998.
39. Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., and Demirbas, M. Short text classification in Twitter to improve information filtering. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, The ACM Press (2010), 841–842.

Heuristic evaluation vs. user testing

Which one to use and when?

DR MARIA PANAGIOTIDI
JAN 20, 2022

Photo by Leon on Unsplash

Heuristic Evaluation and Usability Testing are two different techniques for finding usability problems (Lauesen & Musgrove, 2007). Heuristic Evaluation is done by having usability experts look at an interface, evaluate it against a list of industry standards, and identify problems. In User Testing, potential users try out the interface by performing real tasks and identify problems that impact their experience.

Heuristic Evaluation

In Heuristic Evaluation, experts examine an interface and identify what is good and bad about it. The most popular heuristics used to achieve this are the ones developed by Nielsen and Molich (1990):

1. **Visibility of system status:** The system should always keep users informed about what is going on, through appropriate feedback within a reasonable time.appropriately and promptly.
2. **Match between system and the real world**: The system should speak the users' language, with words, phrases, and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order.
3. **User control and freedom**: Users often choose system functions by mistake and will need a clearly marked "emergency exit" to leave the unwanted state without having to go through an extended dialogue. Support undo and redo.

4. **Consistency and standards:** Users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform conventions.
5. **Error prevention:** Even better than good error messages is a careful design that prevents a problem from occurring in the first place. Either eliminate error-prone conditions or check for them and present users with a confirmation option before they commit to the action.
6. **Recognition rather than recall:** Minimize the user's memory load by making objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate.
7. **Flexibility and efficiency of use:** Accelerators—unseen by the novice user—may often speed up the interaction for the expert user such that the system can cater to both inexperienced and experienced users. Allow users to tailor frequent actions.
8. **Aesthetic and minimalist design:** Dialogues should not contain information that is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility.
9. **Help users recognize, diagnose, and recover from errors:** Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution. regarding errors and solutions.
10. **Help and documentation:** Even though it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, focused on the user's task, list concrete steps to be carried out, and not be too large.

According to Wang & Caldwell (2002): "Using this technique is intuitive, inexpensive, and a quick way of getting a fairly comprehensive usability problem report".

This method, however, has a number of drawbacks. In particular, it detects too many false positives (false alarms) and minor problems that wouldn't bother actual users (Wang & Caldwell, 2002). This can result in designers and developers spending more time and effort fixing less critical issues, which can be costly to companies.

User Testing

User Testing is seen as the best way to identify the real problems that will impact user performance and experience. In this case, actual users are the ones identifying the issues, so there are fewer false positives identified, and the problems found are worth investigating. Unlike Heuristic Evaluation, User testing is more labour intensive as it takes time to prepare the testing materials (e.g., interview guide, tasks), recruit users, and analyse the data. It can also be more expensive to run as participants often require payment for their time.

What does research show?

Wang and Caldwell used Heuristic Evaluation and User Testing to test the usability of a pre-release version of the software they developed. Their major goal was to compare those two methods in terms of efficiency, effectiveness, and cost/benefit.

In their study, Heuristic Evaluation identified 58 unique problems compared to 10 identified in user testing. The problem types and severities were evaluated using Nielsen's five-point severity-rating scale. Score 0 was given to false alarms, 1 to cosmetic problems, 2 to minor usability problems, 3 to major usability problems, and 5 to usability catastrophe. User Testing revealed

mostly major and minor problems, while Heuristic Evaluation has 29.3% of false alarms. The results are presented in the table below.

	Heuristic Evaluation	User Testing
false alarms	29.3%	
cosmetic problems	31%	
minor problems	31%	30%
major problems	8.7%	70%

Percentages of error type per category (data from Wang & Caldwell)

Heuristic Evaluation was cheaper than User Testing in terms of direct cost and time cost; the researchers paid $10.54 for Heuristic Evaluation and $47.30 for User Testing) and it took 15.5 hours to conduct the Heuristic Evaluation including data analysis, while User Testing required 45 hours.

At first glance, Heuristic Evaluation appears to be more appealing as it is less expensive and less time-consuming than User Testing. However, User Testing provided better performance estimates. While Heuristic Evaluation identified more issues a significant number of them were false alarms. On the other hand, User Testing identified mostly major problems. Furthermore, Wang & Caldwell (2002) suggest that some problems associated with learning are hard to be identified by using Heuristic Evaluation alone. For example, in this study, the test users identified two major problems that had not been detected by Heuristic Evaluation.

What does this mean for practitioners?

Heuristic Evaluation appears to be "a useful testing method in the earlier stages of software development". It is a quick and cheaper method that identifies a wide range of problems. Correcting errors is easier and cheaper in earlier design stages compared to later design stages. Early in development, heuristic evaluation has been shown to have a hit rate of around 50% and reports around 50% false problems (Lauesen & Musgrove, 2007). Even though User Testing has a higher cost and requires more time to run, trying to correct false problems is much more costly than any time/money saved by conducting Heuristic Evaluation alone. As a result, using Heuristic Evaluation early on could help the product team to eliminate as many usability problems as soon as possible. User Testing is more appropriate once a functional prototype of the software is available as it can help us detect real usability problems from actual users. User Testing appears to be more effective than Heuristic Evaluation in finding major problems.

Heuristic Evaluation is a **cheaper and quicker** way to detect usability problems. However, it can't replace User Testing, which can provide **more insightful data** from actual users that can have an impact on user experience.

Read More

An Empirical Study of Usability Testing: Heuristic Evaluation Vs. User Testing - Enlie Wang...
*In this study, two different usability-testing methods (Heuristic Evaluation and User Testing) were selected to test...*journals.sagepub.com

Severity Ratings for Usability Problems: Article by Jakob Nielsen
*Severity ratings can be used to allocate the most resources to fix the most serious problems and can also provide a...*www.nngroup.com

User Interface Design: A Software Engineering Perspective
*Description For some software designers the interface is still seen as an add-on after the rest of the program has been...*www.pearson.com

Heuristic Evaluation (Video)
*Summary: Jakob Nielsen explains the heuristic evaluation method, which allows you to judge a user interface design...*www.nngroup.com

Comments

Write a comment...

Customer Complaint Behaviour and Companies' Recovery Initiatives: The Case of the Hello Peter Website

Bhavna Jugwanth
Debbie Vigar-Ellis

Abstract

With technological advancements, consumers now have many different avenues as platforms to communicate or voice their opinions about organisations and service levels. Increased competition and a more vigilant consumer mean that organisations need to keep track of their customers' perceptions of service and product delivery and where necessary, to respond to customer complaints so as to retain, rather than lose customers.

Hello Peter is the world's largest customer service website which was founded by Peter Cheales in 2000 (Arbuckle 2008: paragraph 2). The website allows for consumers to post their complaints and compliments based on their experiences with a particular firm. The organisation is then requested to provide the consumer with a recovery initiative to remedy the failure.

The objectives of the study aimed to assess and categorise the different types of customer complaints on the Hello Peter website, identify the various companies' recovery strategies to these complaints, and where possible to evaluate the effectiveness of these recovery strategies.

Qualitative research techniques were used to gather in-depth data regarding the consumers' reasons for complaining as well as the organisations recovery strategies. The sample size consisted of 1 000 complaints. Inductive Thematic Analysis was used during data analysis to code and create themes for the data collected.

The most common online complaints on the Hello Peter website were regarding delays in company responses, companies promising action and failing to then act, and unhelpful company responses. Common recovery

strategies used by organisations were offering to be in contact with the complainant and acknowledging the customers' complaint. Offering the customer an apology was also a frequently used recovery initiative. From the consumers who responded to the recovery initiative it was found that a relatively low percentage of complainants were impressed with the recovery outcome and process. The findings also indicated that consumers had the most positive responses when an apology was provided as well as a reference number for the consumer to track their complaint. Online recovery strategy recommendations were made.

Keywords: Online complaint behaviour, customer complaints, recovery strategies, Hello Peter website, customer satisfaction with recovery strategies, Inductive Thematic Analysis

Introduction

Consumers are faced with various choices when making a purchase. Given that consumers are generally spoilt for choice, businesses in any industry need to ensure that they offer a high level of customer service in order to secure customer loyalty as well as a strong brand image. Customer service has a direct impact on customer loyalty as consumers' perceptions are difficult to change (Sabharwal, Soch & Kaur 2010: 126).

With technological advancements, consumers now have many different avenues as a platform to communicate or voice their opinions about organisations and service levels. The Hello Peter website, founded by Peter Cheales in the year 2000, is the world's largest customer service website (Arbuckle 2008: paragraph 2). The site allows consumers to report poor customer service, poor product quality, or provide information on good service that they have received. Companies then have the opportunity to remedy the complaint which will be followed by the consumers' response to the service recovery method chosen.

The purpose of the research project was to provide businesses with an idea of what types of failures consumers' complain about on the Hello Peter website and to provide insight into the different service recovery methods and those that are most effective in solving the customers' complaints.

Literature Survey

This section focuses on current literature relating to customer complaint behaviour and service recovery, as well as the role of technology in customer complaint behaviour. This section also provides a brief discussion of the Hello Peter website and its complaint process.

Customer Complaint Behaviour

Complaint behaviour is one possible response to customer dissatisfaction (Crie 2003: 60). Customer complaint behaviour as an action taken by an individual which involves communicating something negative regarding a product or service to either the firm manufacturing or marketing that product or service, or to some third party organisational entity (Ruoh-Nan & Lotz 2009: 107).

Dissatisfaction can lead to a variety of responses (Lovelock & Writz, 2007: 391):

No action: This refers to the circumstance in which the consumer remains loyal despite the problem experienced and the resulting dissatisfaction. This may be due to there being no available alternative (Butelli 2007: paragraph 6).

Private actions*:* These comprise mainly word-of-mouth communication to friends and family. Dissatisfied customers will tell between eight and ten people about bad service they have experienced, and one in every five angry customers will tell 20 people (Hocutt, Bowers & Donavan 2006: 199). The harm caused by dissatisfied customer's talking to friends is minimal however, compared to the harm generated via new technologies such as the Internet and social media. These technologies make it possible for individuals to voice their disappointment with regards to poor service quickly, in large volumes, around the world and in some cases anonymously (Hocutt *et al*. 2006: 199). The customers who choose negative word-of-mouth usually pursue different objectives to those pursuing public actions, such as simply expressing anger and frustration (Butelli 2007: paragraph 6). An exit strategy refers to the condition in which consumers decide not to repurchase or not to utilize the service again. In order for the consumer to

decide to exit or boycott, s/he must have other available alternatives (Butelli 2007: paragraph 6). According to Kurtz and Clow (1998: 54), only 1 out of 26 dissatisfied customers complain to the firm, the remaining 25 show their displeasure by engaging in firm switching behaviour.

***Public actions**:* These include complaint responses made in order to pressure an organisation into rectifying the complaint or offering to refund the buyer (Velázquez, Contri, Saura & Blasco 2006: 495). These public actions may be in the form of voice responses where the complaining behaviour is directed to the parties perceived to be responsible for a dissatisfying experience. Compared to other responses, voice complaints are a direct, confrontational approach to relieving dissatisfaction. By voicing their discontent to a responsible party, consumers may vent their frustration and perhaps more important, get redress for their dissatisfaction (Chan & Wan 2008: 79). Third party responses involve seeking help from outside parties with sanctioning power, such as the media, consumer advocacy groups and legal agencies that are outside the consumer's social circle. By expending relatively significant time and effort on third-party responses, consumers often try to obtain specific remedies for their dissatisfying experience (Chan & Wan 2008: 80).

Usually consumers need to be dissatisfied in order to complain however other factors may be necessary to move the customer from dissatisfaction to complaint. Such factors may be attribution of the cause of dissatisfaction or psycho-sociological characteristics of the individual consumer (Crie 2003: 67). Attribution describes the process of allocating blame. To lead to customer complaint behaviour, the consumer has to identify clearly the party responsible for his/her dissatisfaction during a given consumption incident. Generally, consumers who observe the cause of their dissatisfaction as being stable (the same crisis may happen again) or controllable (consumer feels the organisation could have prevented the cause of dissatisfaction), are more inclined to either leave the organisation or product, or engage in negative word-of-mouth (Crie 2003: 68).

Frustration is a characteristic that can influence the relationship between dissatisfaction and complaint behaviour. The more substantial the frustration the greater the risk of aggressiveness and customer complaint behaviour is (Crie 2003: 68). Frustration arises not only when the objective

assigned to a given behaviour is blocked or interrupted before its fulfilment, but also when the result achieved has a lower level than that sought, or when its realisation requires more resources than the consumer can, wants or expects to spend to reach the desired objective (Crie 2003: 68). Frustration can arise in situations of purchase intention (unavailability of the product or of the brand) or in post purchase situations. Other individual characteristics may also influence complaint behaviour, e.g. loyalty to the brand, product or supplier; the level of quality assessment, the educational level and tastes; the ability to detect quality differences, and perceptions of the cost/profit ratio of the possible actions (Crie 2003: 69).

Recovery Strategies

A recovery initiative refers to the actions an organisation takes in response to a service/product failure (Hocutt *et al*. 2006: 199). Recovery strategies are strategies practiced by an organisation and its employees to return the customer to a state of satisfaction (Nikbin, Ismail, Marimuthu and Jalalkamali 2010: 47). An aim of service recovery is to appease dissatisfied customers through suitable actions in order to reduce potential damage to customer relationships instigated by service failures (Nikbin *et al*. 2010: 47).

While poor complaint management procedures can alienate customers forever, effective recovery strategies offer organisations the opportunity to regain customers through secondary satisfaction or post-complaining satisfaction. One can define customer complaint behaviour as an action taken by an individual who involves communicating something negative regarding a product or service to either the firm manufacturing or marketing that product or service, or to some third party organisational entity (Ruoh-Nan & Lotz 2009: 107). Boshoff and Leong (1998: 24) found that firms accepting responsibility (attribution) for the service failure is the most important factor to customers.

Recovery initiatives illustrate the actions that companies take to counter defects or failures. The most frequent and often used actions are apology, assistance, and/or compensation (Levesque & McDougall 2000: 21). The following section discusses the typical recovery strategies used by organisations to redress failures.

1. Apology: An apology is recommended as a pre-requisite for service recovery. While an apology is better than no apology, an apology alone is relatively ineffective when a customer experiences a failure. Typically, a customer expects some gain for their loss. An apology offers little gain but may be effective when minor problems are encountered (Levesque & McDougall 2000: 21). Through apologies, organisations indicate to complainants that the organisation stands on the same side as the customer and this allows them to work together to solve the problem (Hui & Au 2001: 163). Past findings indicate that providing a respectable explanation can minimise customers' dissatisfaction with poor service experiences. By apologising to complainants, organisations accept responsibility for the problem and express their genuine regret to complainants (Au, Hui & Kwok 2001). The presence or absence of an apology is strongly correlated to customer's perceptions of interactional justice (Wirtz and Matilla 2003: 151).

2. Assistance: Assistance involves taking action to rectify the problem. Assistance is possibly the most effective single recovery strategy, because it can bring the customer back to the original purpose of buying the product/service. It is argued that the service firm has little leeway; it must fix the problem quickly. The gain is fulfilling the basic promise, which may equal the loss from the failure (Levesque & McDougall 2000: 22).

3. Compensation: Compensation involves monetary payment for the inconvenience the customer has experienced and may be required if the failure cannot be fixed. Increasing compensation should lead to greater satisfaction with the recovery strategy (Levesque & McDougall 2000: 7). However, according to Smith, Bolton and Wagner (1999: 369) equity theory suggests that over rewarded customers' may be less satisfied, as they feel distress and guilt about the inequity of the exchange. Thus while consumers want a gain in this loss situation, and increasing the gain through compensation and assistance should improve satisfaction, there may be an upper limit to the gain (Levesque & McDougall 2000: 22). By organisations

offering some kind of compensation, the company is able to decrease the extent of perceived injustice by having an effect on the physical outcome of the complaint. More importantly, compensation is also believed to express a symbolic statement of respect to the complainant and express heartfelt regret of the company. These symbolic meanings are likely to affect perceived fairness of the complaint process (Hui & Au 2001: 163).

Customer Satisfaction with Recovery Strategies

Schoefer (2008: 211) proposes that satisfaction with recovery strategies will be influenced by a customer's perception of i) the way in which s/he was treated during the recovery practice (interactional justice), ii) the means in which conclusions are made and encounters determined (procedural justice), and iii) the perceived result of the complaint (distributive justice).

1. Distributive justice refers to the assignment of tangible resources by the organisation to remedy and reimburse a service failure (Nikbin, *et al*. 2010: 49). Customers may expect various levels of compensation depending on how severely the service failure affects them (Hocutt, *et al*. 2006: 200). In a service recovery effort, tangible compensation will lead to higher perceptions of distributive justice (redress fairness), which in turn will result in higher customer satisfaction and lower negative word-of-mouth intentions (Hocutt, *et al*. 2006: 200). Consumers expect outcomes, or compensation, that corresponds with the level of their dissatisfaction. This compensation can take the form of actual monetary compensation, an apology, future free services, reduced charges, repairs and/or substitutes (Wilson, Zeithaml, Bitner & Gremler 2008: 379).

2. Procedural Justice is the perceived fairness of the process through which results are attained. The perceived fairness of procedural justice is influenced by voice and neutrality. Voice refers to the opportunity that is provided to the consumer to present information about their experience regarding the service failure. Neutrality

occurs when a particular organisation follows a set of processes to redress the situation (Sabharwal, *et al*. 2010: 128). Thus in addition to fair compensation, customers expect fairness in terms of policies, rules and timeliness of the complaint process. Customers want accessibility to the complaint process, and they want things handled quickly, preferably by the first person they interact with (Wilson, *et al*. 2008: 379). A timely response on the part of the front-line employees who are permitted to manage a service failure situation would function as an indication of the suppliers consideration of the consumer's needs (Hocutt *et al*. 2006: 201).

3. Interactional justice is the degree to which customers feel that they have been treated justly while personally interacting with employees of a company during the recovery process. This justice comprises the communication process and treatment of individuals with courtesy, respect and explanation. The capability and enthusiasm of the contact employees to respond and handle service failures can affect the service encounter being remembered as satisfactory of dissatisfactory (Sabharwal *et al*. 2010: 129). Features of this form of justice include interpersonal sensitivity, treating people with dignity and respect, or providing explanations for the events (McColl-Kennedy & Sparks 2011: 253). Interactional justice is the strongest predictor of trust in a supplier as well as overall satisfaction (McColl-Kennedy & Sparks 2011: 253).

Guidelines to Effective Recovery Strategies

The effectiveness of recovery strategies depends on what is done and how it is done (Levesque & McDougall 2000: 21). The following guidelines are an indication as to how organisations can develop a recovery process to ensure customer satisfaction and ultimately customer retention.

- Encourage and track complaints: A critical component of a service recovery strategy is to encourage and track complaints. In many cases it is difficult for the firm to be aware that a service failure has

occurred unless the customer informs the company. A relatively low percentage of customers (5-10%) will complain to an organisation. Firms however can develop strategies to provoke consumers to complain such as developing the mind-set that complaints are good, making complaining easy and being an active listener (Wilson *et al*. 2008: 382). Customers should know where to go and/or who to talk to if they have a complaint. Technological advances have made it possible to provide customers with multiple avenues to complain such as customer call centres, email addresses as well as website feedback forms. Huppertz (2007: 433) states that consumers observe complaining as easier when firms device detailed policies intended to decrease the time and effort necessary to complain. Authorising employees, decreasing the hassle involved in returning goods, as well as providing contact customer service agents make complaining easier. Freephone call centres, emails and pagers are used to facilitate, encourage as well as track complaints.

- Act quickly: Complaining customers want quick responses. Therefore if the company welcomes, even encourages complaints, the firm must be prepared to act on them quickly. Immediate responses require not only systems and procedures that allow quick action but also empowered employees (Wilson, *et al*. 2008: 385). Gordon, McDougall, Terrence and Levesque (1999: 12) found that when a service failure concerning waiting occurred, service recovery strategies (including both assistance and compensation) that were typical of industry practices did not lead to positive future intents towards the service provider. Response speed is one of the main factors of successful service recovery. According to Mattila and Mount (2003: 142), technologically inclined customers seem to have a no tolerance for delayed responses to their electronic complaints. Subsequently these upset customers are able to promptly share their bad experiences with a big number of other consumers through Internet complaint sites; negative word-of-mouth can have a snowball effect on an organisation. Participants who showed a lower level of technology interest were more lenient through a 48 hour period. Cho, Im, Hiltz and Fjermestad (2002: 323) also found that

prompt responses to consumers' complaints are related to repeat purchase intention.

- Take care of problems on the front line: Customers want the persons who hear their complaints to solve their problems whether a complaint is expressed in person, over the telephone or via the internet (Wilson *et al*. 2008: 285). Schoefer (2008: 211) states that it is not the service recovery initiative in itself that produces emotion but rather the manner in which the individual assesses it. Particular emotions and their force are linked to an assessment of the circumstance provoking the emotional response. For example, the polite treatment (i.e. high level of interactional justice) of a customer during service recovery strategies is likely to cause higher levels of positive emotions such as happiness. A rude treatment (i.e. low level of interactional justice) of the consumer, conversely, is likely to increase the possibility of negative emotions such as anger being stimulated (Schoefer 2008: 212).

- Empower employees: Employees must be trained and empowered to solve problems as they occur (Wilson *et al*. 2008: 385). This statement is reinforced by Schoefer (2008: 212) who states that employees should be trained to play their roles in accordance to customer expectations. Schoefer (2008: 212) also states that contact employees should be conscious of the emotional environment of customer complaint management and should be trained to observe it. Employees need training to cultivate emotional capabilities and decision-making expertise. Decision making training can minimise negative emotional responses on customers' perceptions.

- Provide adequate explanations: When customers experience service failures, these individuals try to understand why the failure occurred. Research suggests that when a firm's ability to offer an acceptable outcome is not successful, further dissatisfaction can be reduced if an adequate explanation is provided to the customer (Wilson *et al*. 2008: 387).

- Treat customers fairly: Customers expect to be treated fairly in terms of the outcome they receive, the process by which the service recovery takes place, and the interpersonal treatment they receive (Wilson *et al*. 2008: 387).

Role of Technology in Customer Complaint Behaviour

Consumer complaining is moving from a private to a public sensation. Consumers who once might have voiced their dissatisfaction with a firm to a few family members or friends are now complaining to the first mass media available, to the public World Wide Web (Ward & Ostrom 2006: 220). The evolution of the Internet and its communication potential has given rise to various websites that function as forums for consumers to share their positive or negative experiences when dealing with various organisations (Harrison-Walker 2001: 397). The Internet offers consumers an anonymous and simple available channel for negative word-of-mouth through expressing their viewpoints and/or making complaints available to others. Negative word-of-mouth in the form of consumer criticisms has the potential to taint a brand and sway a potential consumer to search elsewhere for the product (Sparks & Browning 2011: 799).

Not all service failures are expected to lead to online and public actions. Customers usually engage in online public complaining when a service failure is shadowed by failed recoveries (Gregoire, Tripp & Leoux 2009: 19).

Previously, retailers and service providers were unable to redress customer complaints unless the consumer first sought remedy; however this no longer applies (Harrison-Walker 2001: 398). Retailers and service providers who observe complaint forums on the Internet are also in a position to take corrective action (Harrison-Walker 2001: 398). Creating a public forum on the Internet, which can be accessible to a global audience, is a very useful tool for word-of-mouth advertising. The unfortunate side of consumer complaint sites is that consumers seeking information about various organisations will often locate the complaint sites first (Harrison-Walker 2001: 398).

Considering the possible damage that these websites can have on the bottom line of an organisation, many firms such as Volvo and Chase Manhatten are attempting to shield themselves by creating anti-domains, such as chasestinks.com, chaseblows.com etc. (Harrison-Walker 2001: 398). This provides newer firms with an opportunity to block complaint sites before their name is known (Harrison-Walker 2001: 398). Firms that adopt such a defensive stance are attempting to block the consumer's capacity to share their negative incident with others. The damage of dissatisfaction has been acknowledged (Harrison-Walker 2001: 398). At the very least it results in negative word-of-mouth with regard to the inability of the service provider to meet consumer needs, reduced repeat purchases by the dissatisfied consumer and also fewer purchases by new consumers who has been exposed to the negative word-of-mouth (Harrison-Walker 2001: 398).

The key reason for attending to consumer complaints, instead of trying to block them is merely for the reason that it is cheaper in the long run to retain existing customers' satisfied than to spend the marketing monies needed to find new ones. Also, research shows that it costs five times as much to draw a new customer as it does to maintain a current consumer (Harrison-Walker 2001: 399). Whilst in the past an unhappy consumer might tell another 12 to 20 persons about the experience, it appears the reach of complaints expressed on the Internet is virtually endless (Sparks & Browning 2011: 800). As a result, retailers and service providers who are unaware of these consumer complaint forums may unknowingly be losing business because of negative comments made by unsatisfied customers (Harrison-Walker 2001: 398).

According to Butelli (2007: paragraph 15), organisations that do not receive complaints are depriving themselves of the most priceless form of information. It can be seen as 'free' feedback which can provide vital information that is otherwise not available.

The Hello Peter Website

Hello Peter enables consumers to post comments about their experience with a particular company whether it is positive or negative (Arbuckle 2008: 1). The purpose of the website is to improve the service levels of suppliers by

providing a platform for consumers to post company specific constructive criticism as well as compliments (Arbuckle 2008: 2). To date Hello Peter has listed 1 470 companies which are registered with them, and 679 which do not respond to customers complaints. In addition there are 1 321 companies which have been mentioned for the first time in the past 6 weeks and which are still pending and have not become subscribers yet (Hello Peter 2010).

Published customer complaint or complement reports remain on the website for a period for 12 months (Hello Peter 2010). Consumers do not pay to submit a report and can browse other people's reports and search for reports on a particular industry or company (Hello Peter 2010). A company that wishes to subscribe to this service pays an annual subscription fee of R427.50. Additionally, companies are charged an annual response fee according to the number of responses received per annum (Hello Peter 2010). The companies' annual fee includes email notification when a customer report is posted mentioning the particular organisation. In addition to email notification suppliers can choose to have SMS notifications sent to them as well. Each report is accompanied by the customers' name, email and telephone number. The supplier also has the ability to respond to the customers' complaint as well as have access to the customer's rating of the response (Hello Peter 2010).

Research Methodology

Sample Design and Data Collection

The sample population can be described as all individuals who are aware of the Hello Peter website and who utilise the website as a complaint platform. It is therefore difficult to ascertain the population size. A sample size was selected using non-probability sampling due to the type and quality of information needed for the research. Non-probability sampling is described as less complicated and more economical in comparison to probability sampling (Welman, Kruger & Mitchell 2005: 68).

In order to collect data the website was monitored over a two week period (11- 24 July 2011) on four days a week (Monday, Wednesday, Friday and Sunday). Every alternative day was chosen allowing for new complaints

to be posted as well as for suppliers and customers to respond in order to achieve the research objectives. The period of two weeks allowed for conclusive data on the different types of complaints that were on the website as well as, the various recovery strategies that were being utilised.

Based on prior observation of the website it had been noted that there are approximately 500-700 complaints daily. Therefore in order to ascertain a representative sample, the average daily complaints were divided by the days in which research was conducted thus resulting in 125 complaints per day, and ultimately leading to a sample size of 1 000 randomly chosen complaints.

Data Analysis

Inductive Thematic analysis shares several of the actions and principles of content analysis (Marks & Yardley 2004: 57). Thematic analysis is an exploration for themes that develop as being important to the description of the phenomenon (Fereday & Cochrane 2006: 82). An inductive approach means the themes identified are strongly related to the data themselves (Braun & Clarke 2006: 83). A theme refers to a specific pattern found in the data in which one is interested. A further distinction in terms of what represents a theme lies in whether it is drawn from existing theoretical ideas (deductive reasoning) or from the raw information itself(inductive reasoning) (Marks & Yardley 2004: 57). The method involves the identification of themes through vigilant reading and re-reading of the data. It is a form of pattern acknowledgment within the data, where developing themes become the groupings for analysis (Fereday & Cochrane 2006: 82).

Further evaluation was conducted by comparing the recovery strategies to the customers' responses, to identify which recovery strategies are most effective.

Results

Table 5.1 presents product or service failures that have been experienced by consumers in various industries.

Table 1 Types of Complaint Themes

	Theme	Frequency	Percentage
1	Delay in response	459	45.9
2	Promise to do something and didn't	439	43.9
3	Unhelpful	408	40.8
4	Ignored	354	35.4
5	Defective product	227	22.7
6	Bad attitude	209	20.9
7	Rude or impolite	97	9.7

Table 5.1 indicates that the largest category of complaints (45.9%) are due to a delay in response. This theme incorporated statements such as 'to date nothing has happened'. Research suggests that technologically-inclined customers have no tolerance for delayed responses to their electronic complaints, subsequently these upset customers are able to promptly share their bad experiences with other consumers through Internet complaint sites (Mattila & Mount 2003: 142). According to the findings, 43.9% of the complaints were about suppliers who have promised to do something and did not. Statements such as 'Promise to contact you and never do' were used by the complaining customers. This reinforces the literature that states if the promises made by the organisation have not been met, consumers are likely to become dissatisfied (Gordon *et al*. 1999: 8). Unhelpful employees were the 3rd most common complaint theme (40.7%).These consumers used phrases such as 'I called the call centre; no one knew how to help me'. This dissatisfaction is reinforced by Gruber, Reppel, Abosag and Szmigin (2008:132) who state that if the frontline employees are unable to deal with a customer's expectations effectively, the customer is likely to become dissatisfied.

Table 5.2 indicates the various recovery strategies used by suppliers.

Table 2 Recovery Strategy

	Theme	Frequency	Percentage
1	Contact	675	67.5

2	Acknowledgement	479	47.9
3	Apology	470	47
4	Investigate	421	42.1
5	Reference number	228	22.8
6	Compensation	49	4.9

The most frequent supplier response theme (67.5%) was that of suppliers offering to be in contact with the complainant. Studies show that customers being able to voice their complaint produced a significant impact on both perceived fairness of the complaint-handling process and perceived fairness of the complaint result (Hui & Au 2001: 171). A relatively high percentage (47.9%) of suppliers provide acknowledgement to the consumers of their complaint which is reinforced by comments such as 'corrective measures will be put in place'. According to Magnini, Ford, Markowski and Honeycutt (2007:214) trust can be reinforced when partners take action in ways that acknowledge an individuals' specific need and assert their sense of worth. Firms gain trust from the complainants by acknowledging their complaints and providing explanation as to what the firm intends on doing with regard to the complaint. This trust that is gained, can ultimately ensure customer retention.

Offering apologies to complainants was used in 47% of the cases. Supplier comments such as 'please accept my apologies' were found. The literature states that an apology alone is relatively ineffective when a customer experiences a failure (Levesque & McDougall, 2000: 21). Furthermore, Gordon, *et al*. (1999: 12) states that 'doing something' further than apology was significant but not good enough. These statements are an indication that organisations should use an apology as the minimal recovery initiative and not the only strategy. The apology should be combined with other strategies appropriate to the severity of the failure, such as compensation or assistance.

Table 5.3 indicates the various recovery strategies that service providers have utilised and the consumers' response to these efforts.

Table 3 Recovery Strategies and Customer Responses

	Positive responses	Negative responses	Total responses	Total recovery offers
Apology				470
Frequency	155	15	170	
Percentage	32.9	3.2		
Reference Number				227
Frequency	97	4	101	
Percentage	42.6	1.7		
Investigate				421
Frequency	118	14	132	
Percentage	28	3.3		
Contact				675
Frequency	84	23	107	
Percentage	12.5	3.4		
Acknowledgment				479
Frequency	134	28	162	
Percentage	20	4.1		
Compensation				49
Frequency	19	38.8		
Percentage				

According to the findings, providing a customer with a reference number in order to track their complaint results in the highest positive response outcome (42.6%). However, there is no relevant literature to reinforce this finding.

Nearly thirty-three percent of consumers however, had a positive perception when an apology was provided to them by the supplier. According to Levesque and McDougall (2000: 21) an apology alone is relatively ineffective but better than none at all. This again suggests that suppliers

should provide an apology as a minimal recovery initiative and not the only strategy.

Table 5.4 indicates how many consumers problems were solved after the offer of the following recovery strategies.

Table 4 Recovery Strategies Effectiveness on Solving the Problem

	Problem Solved	
Total	114	
	Frequency	Percentage
Apology	90	78.9
Reference Number	52	45.6
Investigate	65	57
Contact	53	46.5
Acknowledge	78	68.4
Compensation	9	7.9

Table 5.4 indicates that the majority of complaints (78.9%) were solved after an apology was offered by the supplier. This is followed by the organisation acknowledging the customers complaint (68.4%). This finding supports Magnini, *et al.*'s (2007: 214) statement that trust can be reinforced when partners take action in ways that acknowledge an individuals' specific need and sense of worth.

Table 5.5 indicates how many consumers are still awaiting further response from the supplier and their problem has not been solved.

Table 5 Recovery Efficiency

	Awaiting Further Response	
Total	17	
	Frequency	Percentage

Apology	11	64.7
Reference Number	3	17.6
Investigate	6	35.3
Contact	8	47.1
Acknowledge	11	64.7
Compensation	0	0

Even though an apology was provided to the customer, 64.7% of these consumers were still awaiting further response from the supplier. According to Mattila and Mount (2003: 142), response speed is one of the main factors of successful service recovery. The authors also state that technologically inclined customers are not tolerant of delays in responses to complaints.

Recommendations

The results of the study indicate that most consumers complain due to their being a delay in response. Delays are no longer tolerated (Mattila and Mount, 2003: 142). Therefore the recommendation is for organisations to reassess their current complaint handling process, and implement controls to alleviate the possibility that customers may experience a delay in response.

According to Hocutt *et al*. (2006:199) the Internet lets people voice their frustrations regarding poor service quickly, in great volume, around the world, and anonymously. Therefore it is recommended that companies develop their own websites to deal with customer complaints, in this way allowing the organisation more control of what information is shared with vast numbers of potential customers. This is reinforced by the actions of Volvo who have created their own anti-domain (Harrison-Walker 2001: 398).

Limitations

Complaints on the Hello Peter website with customer updates are not done immediately after the recovery initiative has been executed, therefore the findings may lack representation with regard to that objective as not all

complaints had customer responses to the recovery initiative. It did however enable one to get an idea of the customers' opinions of the recovery initiatives. As with all qualitative data analysis, a fair amount of subjectivity occurs. The authors however, attempted to reduce this subjectivity by comparing responses between themselves as well as to the literature on complaint behaviour and recovery strategies.

Conclusion

The Hello Peter website has revolutionised the way customers complain about their experiences, as well as affected the way organisations attempt to remedy these failures. The purpose of the study was to determine what customers were complaining about, what recovery initiatives companies used and where possible, the effectiveness of the recovery strategies. Customers most frequently complained about experiencing a delay in response, organisations promise to do something but do not. The most common recovery initiatives were organisations offering to be in contact with the complainant as well as acknowledgement of the complaint.

The study has provided information and suggestions in order for companies to improve their current online complaint handling strategies as well as develop insight into the most effective service recovery initiatives from the customers' perspective.

References

Arbuckle, A 2008. Tired of Poor Service? Say Hello to Peter. Available at: http://www.witness.co.za/index.php?showcontent&global%5B_id%5D=6149. (Accessed on October 2012.)

Au, K, M Hui & L Kwok 2001. Who Should be Responsible? Effects of Voice and Compensation on Responsibility Attribution, Perceived Justice, and Post-complaint Behaviors across Cultures. *International Journal of Conflict Management* 12,4: 350.

Boshoff, C & J Leong 1998. Empowerment, Attribution and Apologising as Dimensions of Service Recovery: An Experimental Study. *International Journal of Service Industry Management* 9,1:24 - 47.

Braun, V & V Clarke, V 2006. Using Thematic Analysis in Psychology.

Qualitative Research in Psychology 3,2:77 - 101.

Butelli, S 2007. Consumer Complaint Behaviour (CCB): A Literature Review. Available at: http://dspace-unipr.cilea.it/bitstream/1889/1178/1/Butelli%2520Literature%2520review.pdf. (Accessed on October 2012.)

Chan, H & L Wan 2008.Consumer Responses to Service Failures: A Resource Preference Model of Cultural Influences. *Journal of International Marketing* 16,1:72 - 97.

Cho, Y, I Im, R Hiltz & J Fjermestad 2002. The Effects of Post-Purchase Evaluation Factors on Online vs. Offline Customer Complaining Behavior: Implications for Customer Loyalty. *Advances in Consumer Research* 29,1:318 - 326.

Crie, D 2003. Consumers' Complaint Behaviour. Taxonomy, Typology and Determinants: Towards a Unified Ontology. *Journal of Database Marketing & Customer Strategy Management* 11,1:60 - 79.

Fereday, J & EM Cochrane 2006. Demonstrating Rigor Using Thematic Analysis: A Hybrid Approach of Inductive and Deductive Coding and Theme Development. *International Journal of Qualitative Methods* 5,1:80 - 92.

Gordon, HG, G McDougall, J Terrence & T Levesque 1999. Waiting for Service: The Effectiveness of Recovery Strategies. *International Journal of Contemporary Hospitality Management* 11,1:6 - 15 .

Gregoire, Y, TM Tripp & R Leoux 2009. When Customer Love Turns into Lasting Hate: The Effects of Relationship Strength and Time on Customer Revenge and Avoidance. *Journal of Marketing*73,6:18 - 32.

Gruber, T, A Reppel, I Abosag & I Szmigin 2008. *Preferred Attributes and Qualities of Effective Customer Contact Employees during Face-to-Face Complaint Handling Encounters: A Cross-National Comparison Study.* St. Petersburg, FL, USA: Society for Marketing Advances Conference.

Harrison-Walker, L 2001. E-complaining: A Content Analysis of an Internet Complaint Forum. *Journal of Services Marketing* 15,4/5: 397 - 412.

Hello Peter 2010. Home. Available at: http://www.hellopeter.com/. (Accessed on October 2012.)

Hocutt, M, M Bowers, & D Donavan 2006. The Art of Service Recovery: Fact or Fiction? *Journal of Services Marketing* 20,3:199 - 207.

Hui, M & K Au 2001. Justice Perceptions of Complaints-handling: A Cross-cultural Comparison between PRC and Canadian Customers. *Journal of*

Business Research 52,2:161 - 173.

Huppertz, JW 2007. Firms' Complaint Handling Policies and Consumer Complaint Voicing. *Journal of Consumer Marketing* 24,7:428 - 437.

Kurtz, LD & EK Clow 1998. *Services Marketing*. New York: Wiley.

Levesque, T & G McDougall 2000. Service Problems and Recovery Strategies: An Experiment. *Canadian Journal of Administrative Sciences* 17,1:20 - 37.

Lovelock CH & J Wirtz 2007. *Services Marketing: People, Technology, Strategy. 6th Edition.* New Jersey: Pearson Prentice Hall.

Magnini, V, J Ford, E Markowski & J Honeycutt 2007. The Service Recovery Paradox: Justifiable Theory or Smoldering Myth? *Journal of Services Marketing* 21,3:213 - 224.

Marks, D & L Yardley 2004. *Research Methods for Clinical and Health Psychology*. Available at: http://books.google.co.za/books?hl=en&lr= &id=SHiUvmKzuFwC&oi=fnd&pg=PA56&dq=thematic+inductive+analysis&ots=JmRTvc-y5R&sig=crrGSqfGnDnLAJVf0FJfXwcF__c#v= onepage&q =thematic%20inductive%20analysis&f=false. (Accessed on October 2012.)

Mattila, A & D Mount 2006. The Impact of Timeliness on Complaint Satisfaction in the Context of Call-Centers. *Journal of Hospitality & Leisure Marketing* 14,3:5 - 16.

McColl-Kennedy, JR & BA Sparks 2011. Application of Fairness Theory to Service Failures and Service Recovery. *Journal of Service Research* 5,3:251 - 266.

Nikbin, D, I Ismail, M Marimuthu, & M Jalalkamali 2010. Perceived Justice in Service Recovery and Recovery Satisfaction: The Moderating Role of Corporate Image. *International Journal of Marketing Studies* 2,2:47 - 56.

Ruoh-Nan, Y & S Lotz 2009.Taxonomy of the Influence of Other Customers in Consumer Complaint Behavior: A Social-psychological Perspective. *Journal of Consumer Satisfaction, Dissatisfaction & Complaining Behavior* 22:107 - 126.

Sabharwal, N, H Soch & H Kaur 2010. Are we Satisfied with Incompetent Services? A Scale Development Approach for Service Recovery. *Journal of Services Research* 10,1: 125 - 142.

Schoefer, K 2008. The Role of Cognition and Affect in the Formation of Customer Satisfaction Judgements Concerning Service Recovery

Encounters. *Journal of Consumer Behaviour* 7,3:210 - 221.

Smith, KA, NR Bolton & J Wagner 1999. A Model of Customer Satisfaction with Service Encounters Involving Failure and Recovery. *Journal of Marketing Research* 36,3:356 - 372.

Sparks, BA & V Browning 2011. Complaining in Cyberspace: The Motives and Forms of Hotel Guests' Complaints Online. *Journal of Hospitality Marketing & Management* 19,7:97 - 818.

Velázquez, B, G Contrí, I Saura, & M Blasco 2006.Antecedents to Complaint Behaviour in the Context of Restaurant Goers. *International Review of Retail, Distribution & Consumer Research* 16,5:493 - 517.

Ward, JC & AL Ostram 2006. Complaining to the Masses: The Role of Protest Framing in Customer-Created Complaint Web Sites. *Journal of Consumer Research* 33,2:220 - 230.

Welman, C, F Kruger & B Mitchel 2005. *Research Methodology*. Cape Town: Oxford.

Wilson, A, VA Zeithaml, MJ Bitner &DD Gremler 2008. *Services Marketing Integrating Customer Focus across the Firm. 1st European Edition.* New York: McGraw-Hill Companies.

Wirtz, J & AS Matilla 2003. Consumer Responses to Compensation, Speed of Recovery and Apology after a Service Failure. *International Journal of Service Industry Management* 15,2:150 - 166.

Bhavna Jugwanth
School of Management, IT & Governance
University of KwaZulu-Natal
South Africa

Debbie Vigar-Ellis
School of Management, IT & Governance
University of KwaZulu-Natal
South Africa
VigarD@ukzn.ac.za